

General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.
- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.
- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.
- This document is paginated as submitted by the original source.
- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

R-635

ROBOT VISION

April 1973

by

LOUIS L. SUTRO
Charles Stark Draper Laboratory

JEROME B. LERMAN
Artificial Intelligence Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

(NASA-CR-133458) ROBOT VISION Final
report (Massachusetts Inst. of Tech.)
78 p HC \$6.00

N73-27936

CSCD 16P

Unclas
G3/4 15278

Based on a paper presented to
The First National Conference on Remote Manned Systems
in September 1972

**CHARLES STARK DRAPER
LABORATORY**

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

CAMBRIDGE, MASSACHUSETTS, 02139



R - 635

ROBOT VISION

by

LOUIS L. SUTRO
ASSISTANT DIRECTOR

The Charles Stark Draper Laboratory
A DIVISION OF MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Cambridge, Massachusetts

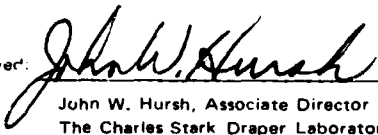
and

JEROME B. LERMAN
Artificial Intelligence Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Cambridge, Massachusetts

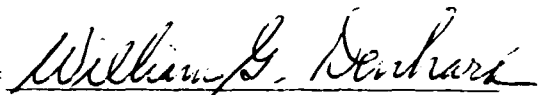
April 1973

*Based on a paper presented to
The First National Conference on Remote Manned Systems
in September 1972*

Approved:


John W. Hursh, Associate Director
The Charles Stark Draper Laboratory

Approved:


William G. Denhard, Associate Director
The Charles Stark Draper Laboratory

ACKNOWLEDGEMENT

The contributions of the following are gratefully acknowledged: Frederick Zeise, Richard Warren, Charles Sigwart and Dr. Roberto Moreno-Diaz in modelling neurophysiology; Larry Baxter, Jerome Krasner, and David Tweed in the design of electronics; Robert Magee in the design of optics; Joseph Convers and Benjamin Smith in mechanical design; Daniel Moulton, James Bever and John Hatfield in programming; and John Hursh in administration. Special appreciation is due the late Dr. Warren McCulloch, Prof. Jerome Lettvin, Dr. Bela Julesz and Prof. Whitman Richards, for describing vertebrate vision in language that the authors and the above could understand and attempt to model.

The work reported here was sponsored by the National Aeronautics and Space Administration, Headquarters, Office of Space Sciences and Applications, through Contract NSR 22-009-138.

ABSTRACT

Here is described the operation of a system built both to model the vision of primate animals, including man, and serve as a pre-prototype of a possible object recognition system. It was employed in a series of experiments to determine the practicability of matching left and right images of a scene to determine the range and form of objects.

The experiments started with computer-generated random-dot stereograms as inputs and progressed through random-square stereograms to a real scene. The major problems were the elimination of spurious matches, between the left and right views, and the interpretation of ambiguous regions, on the left side of an object that can be viewed only by the left camera, and on the right side of an object that can be viewed only by the right camera.

Rules were developed for eliminating spurious matches in the progressively more difficult objects. An arbitrary rule was developed for interpretation of ambiguous regions.

In the experiments reported, comparison of left and right views was performed in terms of gray values, but the comparison could be made in terms of edges. An economical method of detecting edges was demonstrated.

Stereo camera assemblies were designed and one of them built to permit the cameras to converge and together pitch, roll and yaw. A second stereo TV camera assembly has been built which has as its only moving parts two mirrors and a means of focussing. The above experiments were performed, before either of these camera assemblies was available, by exposing a single-view camera in two positions.

We show that a scene on Mars, reported to earth in terms of its features, can be reconstructed on earth.

Perhaps the two most useful results were (1) development of

the concept of a match space where the detected three-dimensional properties of a scene can be plotted and then examined for their form, and (2) the conclusion that a stereo TV camera is needed which will acquire both central and peripheral stereo pairs of images.

TABLE OF CONTENTS

Chapter	Title	Page
I	INTRODUCTION	1
A.	The Main Objective	1
B.	A Supporting Objective	1
C.	Test Systems	2
D.	Approach	4
E.	Other Supporting Objectives	4
F.	How This Paper Differs from First Version	4
II	MEANS OF AUTOMATICALLY DETERMINING THE RANGE OF SMALL AREAS OF A SCENE	7
A.	Objective	7
B.	Opto-Electro-Mechanical Strategies	7
C.	Automatic Comparison of One Scan Line of Left with One Scan Line of Right TV Image	9
D.	Methods of Mapping Binocular Space into Match Space	12
E.	How Matches Are Made and Viewed	14
F.	Use of Model in Match Space	18
G.	Geometry of Binocular and Match Spaces	19
H.	Rules for Processing Random-dot Stereograms	22
I.	How Should Ranges in a Scene Be Presented?	27
J.	Eliminating Areas of Spurious Matches	27
K.	Random-Square Stereograms	32
L.	Processing of a Real Scene	32
M.	Determining Form	33
N.	Range Accuracy	38
III	COMPUTATION TO EXTRACT OTHER FEATURES	41
A1	Square-Wave Frequency Response of Camera	41

Chapter	Title	Page
A2.	Computation of Edges	41a
B.	Formation of Line Drawing	43
C.	Hardware to Detect Edges	46
D.	Computation of Reflecting Properties	48
IV	RECONSTRUCTING THE APPEARANCE OF A SCENE	49
V	STEREO TV CAMERAS	51
VI	SUMMARY AND CONCLUSION	57
APPENDIX A		
	EQUATIONS FOR RANGE AND UNCERTAINTY IN RANGE	61
A. 1	The Geometry of Stereo TV Optics with Parallel Axes	61
A. 2	Derivation of Equation of Range Uncertainty for Stereo TV Cameras with Parallel Axes	62
REFERENCES		65

LIST OF ILLUSTRATIONS

Fig. No.	Brief Title	Page
0	Equipment for simulating light-weight, low-power hardware: Camera-computer chain and binocular, or stereo, TV camera.	3
1	Mars-like scene.	5
2	Two strategies of visual processing.	6
3	Diagram of computation to form a model in match space of an object in binocular space.	11
4	Binocular space divided into quadrilaterals of uncertainty by rays drawn from pixels in left and right images.	13
5	Computation of first-stage matches at the right ends of three lines of the x' -d plane pictured in Fig. 3.	15
6	(a) Geometry of Fig. 3, when object viewed is between parallel optical axes. (b) Match space corresponding to the binocular space in (a).	20
7	Geometry of Fig. 3 when the object viewed is at the left of both optic axes.	21
8	Why the range of the parts of the background, at the left and right sides of an object, is ambiguous.	23
9	A random-dot stereogram depicting a square floating before a background.	24
10	An x' -d section of the m-space generated from the stereogram in Fig. 9.	26
11	x' -d sections of m-space showing, in (a), ambiguous regions on the left and right sides of the model of the object and, in (b), how both the image of the object and the background behind it are modelled.	26

Fig. No.	Brief Title	Page
12	x'-d sections of three match spaces formed from the same stereogram.	29
13	Diagram showing the functions of the three simulation programs, STEREO, EXPER and FUSER.	30
14	Random-block stereogram of a square floating before a background.	33
15	x'-d section of model in m-space generated from the stereogram in Fig. 14.	33
16	Stereo images of the Mars-like scene of Fig. 1, acquired by television camera, digitized, then displayed one at a time on an oscilloscope.	35
17	x'-d section of the match space formed from the stereogram of Fig. 16.	35
18	Range map of one match space of the stereogram in Fig. 16.	37
19	Test pattern and amplitude response of TV camera.	40 and 41a
20	Operations performed on each 7 x 7 pixel array of a digitized image to detect edges.	41b
21	Negative of a display of the result of performing the operations of Fig. 20 on the images of Fig. 16.	44
22	Edges of Fig. 21 after thinning.	44
23	Detection of edge and thinning of edge.	45
24	Insertion of means of detecting edges between the stereo TV camera assembly and match space.	47
25	Effect of a source of collimated light such as sunlight.	50
26	Reconstruction of the appearance of an object from its shape, its reflectance and the sources of light.	50

Fig. No.	Brief Title	Page
27	Type C3b stereo TV camera assembly.	52
28	Type D1 stereo TV camera assembly.	53
29	Type E1 stereo TV camera assembly.	53
30	Diagram of images erected on the face of each vidicon in the E1 camera assembly.	55
31	Possible configuration of a Mars rover.	56
A-1	Geometry of parallel optics for range finding.	63

LIST OF TABLES

Table No.		Page
1	VALUES OF THE MATCH VARIABLES	17
2	TEST CONDITIONS FOR EXAMPLE OF II L	36

I. INTRODUCTION

A. The Main Objective

The main objective of the work reported here was to develop automatic means of classifying three-dimensional objects. The problem requiring solution at the start of this work was how to guide an automatic vehicle (robot) in the exploration of the surface of Mars, recognize objects and report these findings to earth. Problems of a similar nature have since arisen. One is the automatic classification of plankton when viewed under a binocular microscope. Another is the automatic classification of objects to be machined, assembled or otherwise manipulated in the course of manufacture.

While classification is the goal, we have found the word "recognition" easier to use. Thus, "automatic recognition of objects" is the goal most often mentioned here.

B. A Supporting Objective

Since the only systems able to recognize three-dimensional objects are animals, an objective that was pursued, supporting the main objective, was to model animal vision. The first efforts in this direction were to devise models of the vision of a lower vertebrate, namely, the frog. The reasons for giving attention to this animal were, first, that it performs recognition within its eyeball, and, second, that the neurophysiology of the frog's eye is well understood. For example, a moving insect is recognized there and reported via the frog's optic nerve to its brain. Thus, by modelling a frog's eye, one devises an operating model of a recognition system. Reports on this part of our effort describe a first crude model of the frog's eye (Ref. 1), a fine grain model of the bug-detector cell in the frog's eye (Ref. 2), a more rigorous model of this bug detector (Refs. 3-5),

and a first description of the shift register scheme (Ref. 6).

McCulloch described animal nervous systems as multiple loops of information flow, with computation in every loop (Ref. 7). Except for the frog (Ref. 8), however, it was not possible to describe these systems in sufficient detail to enable immediate progress to a useful design. What was needed was an operating model which could be shaped by a series of small changes. Particularly needed was a test system which would permit modelling the binocular vision of primates, including man.

C. Test Systems

Two test systems were built, both to model the vision of animals, including man, and serve as pre-prototypes of possible object recognition systems. The first system is described in Ref. 9. The second is pictured in Fig. 0. The TV camera of this second system was usually aimed at the simulated Mars scene (Fig. 1), which consisted of rolling terrain, made of papier-mâché on a 4 ft. by 8 ft. piece of plywood, and a painted backdrop, all created by Dustin Thomas. Lighting was usually unidirectional from the right. The Type C3 camera was occasionally aimed out the window to test the ability of the system to determine the range of objects on the roofs of adjoining buildings. The edge of a building on the Boston skyline was imaged to serve as an infinity point in the adjustment of the mirrors of the camera assembly.

The TV camera, the central element in stereo TV, generated a TV image displayed 30 times a second on the monitor. The TV camera was also scanned slowly along vertical lines, under computer control, to acquire an image for processing. Any image stored either on magnetic tape or in core memory can be displayed and photographed on the oscilloscope, the intensity of which can be modulated to display gray values.

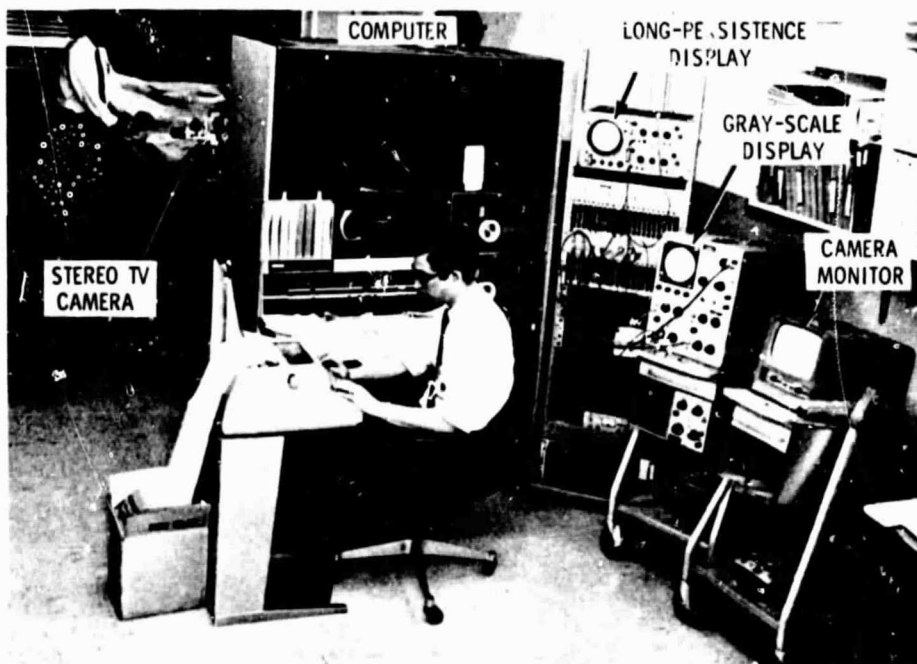
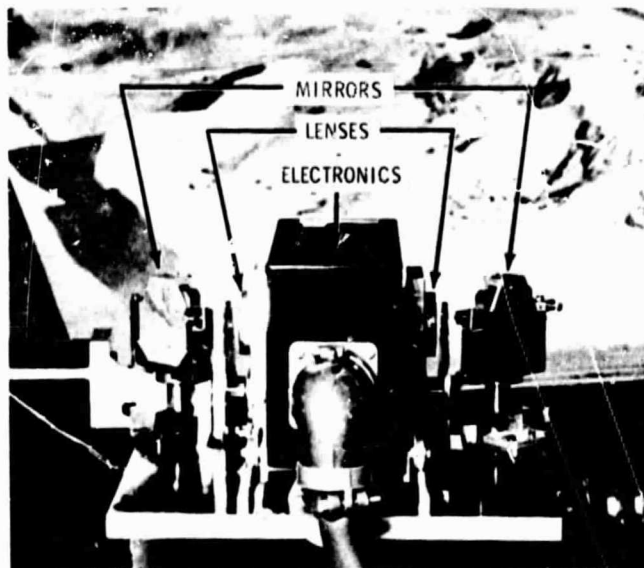


Fig. 0. Equipment for simulating light-weight, low-power hardware
 (above) Binocular, or stereo, TV camera Type C3a
 (below) Camera-computer chain for simulating robot vision

D. Approach

Once this equipment was completed, we turned our attention to making it work. There were two examples before us. One was the modelling of the vertebrate visual system, one cell type at a time, progressing from the retina through the lateral geniculate nucleus to the visual cortex. Fukushima demonstrated that this was possible (Refs. 10, 11).

The other approach was that of Julesz, in which he had employed a computer to compare the unprocessed left and right images of a stereogram to extract range data, thus assuring that the structure of the scene would be acquired automatically. We decided to follow the latter route. Having done that, we now find that it is often desirable to introduce another stage of computation between the acquisition of left and right images of a scene and comparison of these images. It might appear that we will thus arrive at the same result as if we had taken the first approach. In fact, however, by providing for the structure first we have sought and found economies in data handling that we might not have found in taking the first approach.

E. Other Supporting Objectives

To move toward the main objective stated in A above, other supporting objectives had to be pursued; namely, detection of edges with a minimum amount of hardware, reconstruction of the appearance of a scene from detected features, design of stereo TV cameras and design of the mounting of one such camera on a rover.

F. How This Paper Differs from First Version

The first printed version of this paper is Ref. 12. It is revised here to provide a more complete introduction, to clarify the

description of the simulation program EXPER in II J, show in II N how range accuracy can be increased over that in the example of II L, propose another method of stereo processing in II I, expand III to include examples of the detection of spatial frequency, describe in more detail the development of stereo TV cameras in V, and add comments on how this work models primate vision. The only new illustration in Sections I and II is Fig. 0, placed ahead of all of the previously-numbered illustrations. With the addition of Fig. 19 to Section III, all of the previously-numbered figure numbers, from 19 to 29, move up one. With the addition of Fig. 30, the final figure number move up two. Because many more references have been added, the reference numbers differ here from the first printed version.

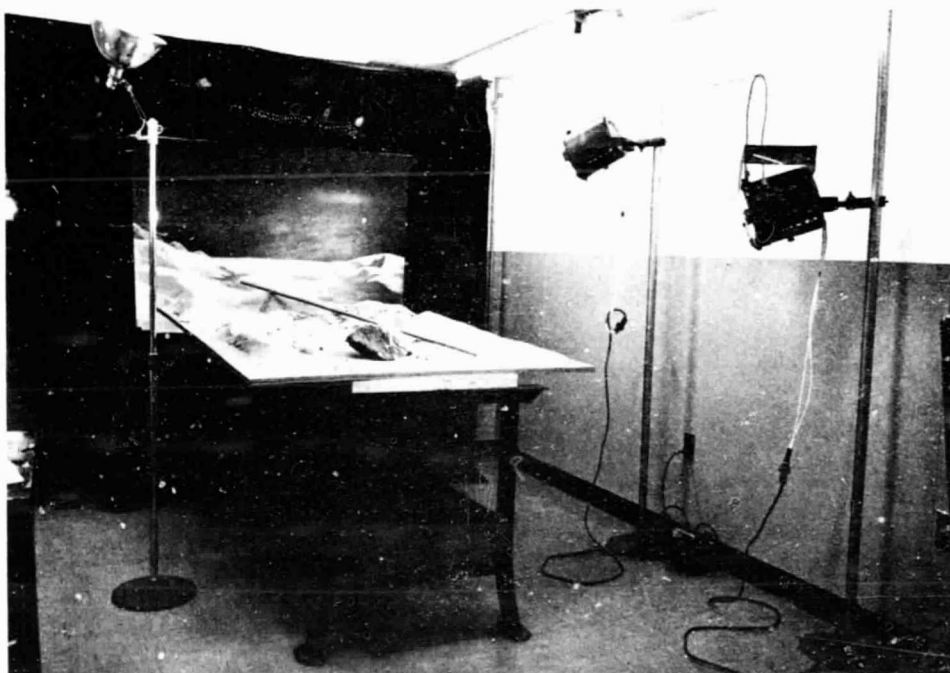


Fig. 1. Mars-like scene.

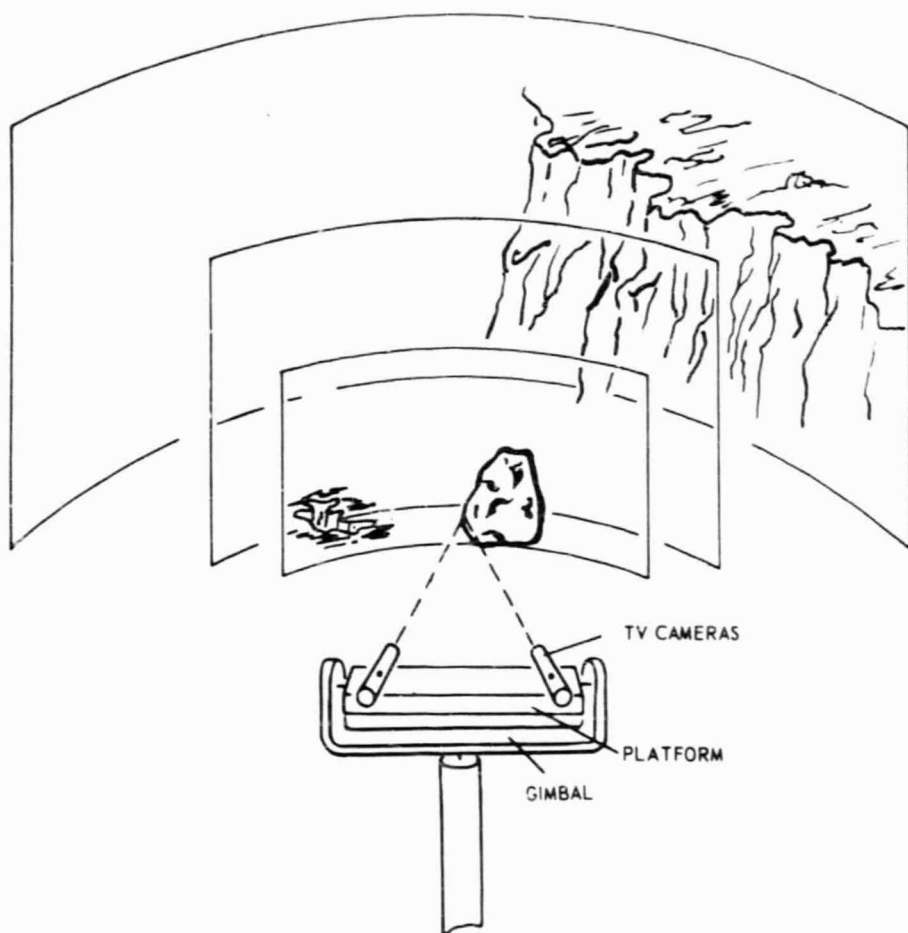


Fig. 2. Two strategies of visual processing: Locating, represented by rectangles, and identifying, represented by dashed lines.

II. MEANS OF AUTOMATICALLY DETERMINING THE RANGE OF SMALL AREAS OF A SCENE

A. Objective

The immediate objective of the work reported in this section was automatic determination of the range of small areas on the surfaces of objects. A further objective was to employ this range information to automatically determine the form of those objects.

The method employed in pursuing the first objective consisted of comparing the gray values of picture element (pixels) in a left image of the field of view with gray values of picture elements in a right image. Because this appeared to be the simplest possible way of comparing left and right images it enabled us to define concepts basic for future work, such as match space, spurious matches and ambiguities. For that future work, comparison of other variables in the left and right images is described in III.

Two automatic means of determining range compete for our attention. One uses a laser beam that is deflected by a mechanically-driven mirror under computer control. The other uses a stereo TV camera which is both controlled and interpreted by computer. Because our long-term goal is automatic recognition of objects characterized by edges and lines (features) which a laser may not be able to detect, we pursued the development of stereo-TV-computing. Laser range-finding can supplement stereo-TV-computing to determine to greater accuracy the range of objects selected by stereo-TV-computing.

B. Opto-Electro-Mechanical Strategies

In bringing a stereo pair of images onto the face of a camera tube or tubes, several opto-electro-mechanical strategies are possible. Figure 2 illustrates two that can be employed consecutively.

The first works on a coarse scale and is called "locating". The second works on a fine scale (high resolution) and is called "identifying" (Ref. 13). Following the first strategy each camera subtends the wide view represented by the rectangles to discover the rock, the hole and the cliff. Following the second strategy the two cameras investigate with higher resolution optics details along the edge of the rock.

As part of the first strategy, the cameras of Fig. 2 are shown mounted on a table that tilts within one gimbal and turns on another. Each camera is also supported by a gimbal on the platform, represented by a black dot on the camera case. The range of each object can be computed from the angles formed by the camera axes and the distances between the cameras, when both cameras are centered on the object. The first strategy has been pursued in the design of the gimballed mirrors and gimballed cameras described in Section V.

The second strategy is to identify the features and, from the position of those features in three-dimensional space, the form of objects. The second strategy has been pursued in the work described in Sections II and III of this paper. In Section II the features are gray levels formed from the pattern of luminances in the scene. In Section III they are edges.

The second strategy consists of two sub-strategies. The first substrategy, called "stereopsis", requires two views and yields in man "the experience of relative depth only" (Ref. 14). The method of comparing two views, on the other hand, that we have designed, yields absolute depth. The second substrategy, called "cognitive processing", requires the storage of features and the relations between features so that these can be compared to features and their relations in the image. This second substrategy is not employed in the examples presented in this paper. How it could be employed is considered in II H.

A third strategy, while not optical or mechanical, is electronic

in the sense of being computational. It can determine the class of thing the stereo-TV computer looks for. This strategy was described first in "Assembly of Computers to Command and Control a Robot" (Ref. 15), and is being described in more detail in Refs. 16 and 17. In those reports the word "robot" is used, as McCulloch used it, to mean an animal or a machine (Ref. 18).

C. Automatic Comparison of One Scan Line of Left with One Scan Line of Right TV Image

Both of the first two strategies require a method of comparing left and right images. The method about to be described stems from the work of Bela Julesz (Ref. 19), and was developed into its present form by the second author (Ref. 20).

Fig. 3 shows at the left two TV cameras whose parallel optical axes extend into a space with coordinates x , y and z (measured in meters). At the lower right is a three-dimensional structure where a model is formed of the scene at left. It has the dimensions, shown in the lower right corner: x' and y' in pixels, d in integral values of disparity. We call the structure "match space". Disparity is related to range by Eq. (1) in II G.

In our simulations of proposed hardware, match space is only 36 values of disparity deep, from $d=0$, corresponding to infinity, to $d=35$. Only one $x'-d$ plane is shown in the match space (m-space) of Fig. 3. Note that the black squares in this plane approximate the shape of a section of the rock at left. Each black square represents a 1 in the computer memory.

The information in this match plane comes from the scan lines on the face of the left and right camera tubes. The left scan line is an image of a V-shaped area projecting from the left camera lens into space, the right scan line the image of a similar V-shaped area. Where the two V-shaped fields overlap is the binocular field of the

cameras.

As the electron beam in the left camera tube starts to sweep across the line pictured on the face of that tube, the voltage of the camera's output is converted to one of 2^5 or 32 levels, which we call "gray levels". Expressed as a five bit word, this gray level enters the first column of the left-image shift register. When the electron beam advances to the next pixel, the first column is shifted to the right and a new column takes its place. This process continues until the left-image shift register is full.

After 36 five bit words have been formed and shifted, as described, the same process begins in the right camera tube and right-image shift register. Thus, when the left image reaches the end of its shift register, the right is only 36 positions behind. Thereafter, the right marches past the left and the two are compared after each shift.

The number of pixels by which the image of a point in the right image falls short of overlapping the image of the same point in the left image is called the disparity of that point. In Fig. 3, the effect is shown of comparing the left and right images when the disparity between left and right images of a point is 35, 34, 33, 32, 31 and 30. For example, when the disparity is 35, a spurious match is formed due to the fact that images with the same gray level are not necessarily images of the same point in the scene. When the disparity is 34, two matches are made, one for each side of the rock. The process of shifting and comparing continues until, at $d=30$, the edges of the rock have been mapped.

That comparisons are made for only 36 values of disparity is due to the size of the memory in the computer used for the simulations. Only one form of an $x'-d$ plane is shown in Fig. 3, namely, one in which successive lines of matches are "justified" to the right, as viewed from the direction of the camera. (The word "justify" is a printer's term which means to line up lines of type evenly.)

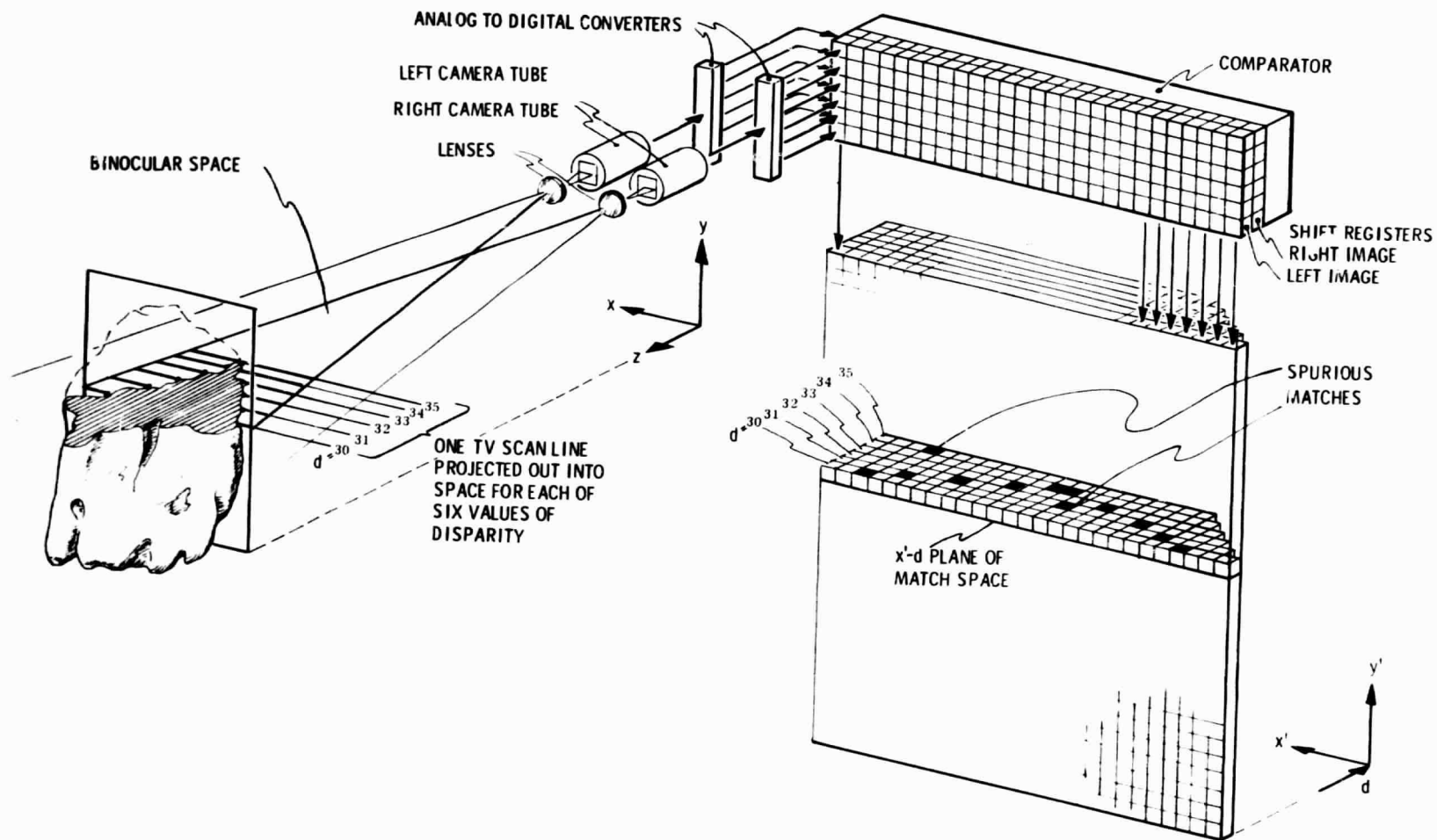


Fig. 3. Diagram of computation to form a model in match space of an object in binocular space.

D. Methods of Mapping Binocular Space into Match Space

Fig. 4 shows a plane of binocular space divided into quadrilaterals of minimum uncertainty, determined by the division of each scan line into pixels. We say "minimum" because the number of quadrilaterals will be as few as illustrated only if, (1) a camera tube is employed, of which the maximum resolution in TV lines approximates the division into pixels shown here, and (2) the spatial frequency of the high-contrast detail in the scene is that which leads to this resolution.

The uncertainty pictured in Fig. 4 is due to the choice of focal length of the lens and to the characteristics of the camera tube. This uncertainty is optoelectronic. The uncertainty due to positioning of the camera is electromechanical. The need to add the two uncertainties can be obviated, at least during a fixation, by rigidly attaching the optoelectronics for strategy 1 to the optoelectronics for strategy 2. The Type E1 stereo TV camera, described in V, is designed this way.

Fig. 4 shows the effect of projecting the pixels on the face of the two camera tubes out into binocular space. Because we see them here only in plan view, we call the intersection of two pixel rays from the left a "quadrilateral of uncertainty". Actually it is a polyhedron of uncertainty (Ref. 21).

Note that the number of quadrilaterals formed by intersecting rays increases as z increases.

When the lines of matches formed in match space are justified to the left, as shown in the lower right corner of Fig. 4, a right-camera view is formed. This can be verified by following, first, the left ray of the right camera which can be seen to form a continuous succession of quadrilaterals with rays from the left camera. Next follow the succession of rays from the right camera which intersect a single ray from the left camera in a manner that leads to the jagged

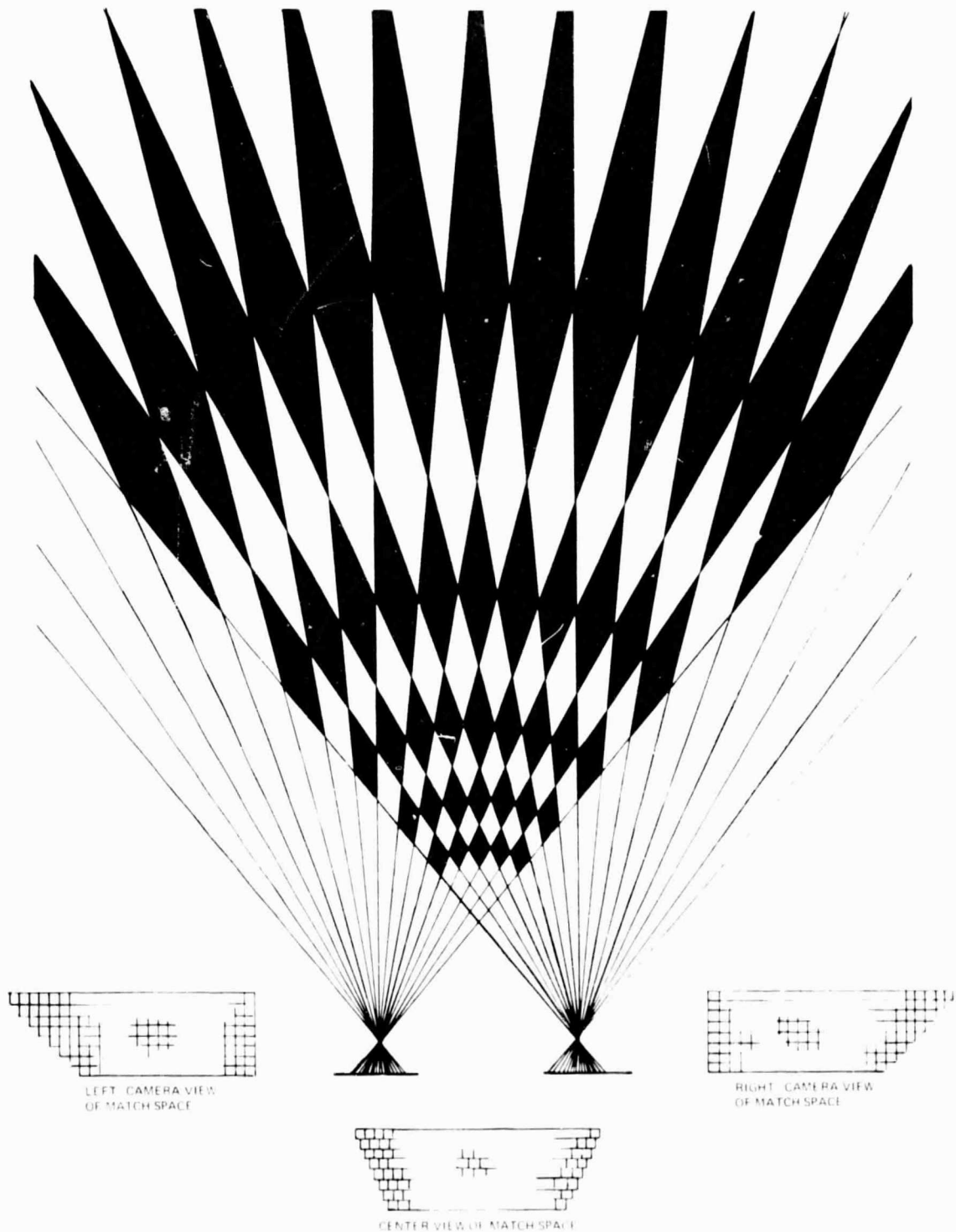


Fig. 4. Binocular space divided into quadrilaterals of uncertainty by rays drawn from pixels in left and right images. Along any ray only alternate quadrilaterals of uncertainty are shaded.

right edge of the right-camera view.

A left-camera view and a center view are also presented in Fig. 4. The left- and right-camera views are needed for the method of eliminating spurious matches of areas presented in II J. In hardware, only one model of matches will need to be formed of binocular space. The two views can be obtained by two sets of inter-wiring.

E. How Matches Are Made and Viewed

Fig. 5 shows how first-stage matches* are made in the system of Fig. 3. Fig. 5 pictures, from above, the right ends of the two shift registers at three different positions of one scan line of the right image. At the first position, where disparity equals 32, two matches are made, one of them spurious. In the second position, where $d=31$, another match is made and in the third position where $d=30$, another.

Exact ($\epsilon=0$) first-stage matches such as these can usually be made only between computer-generated images. Between images of a real scene, tolerances are needed at both first and second stages of match, to allow for noise in the electronics or noise in the scene. By the latter we mean, for example, unequal reflection of light from the same spot in the scene to the different viewpoints of the two cameras.

The tolerance at the first stage of match is the allowed difference between the gray value of one pixel in the left image that is considered matched to a gray value of one pixel in the right image. We call this tolerance ϵ and assign it the values shown in the second column of Table 1.

The tolerance at the second stage of match is in the form of a threshold and accompanies our requirement that an $N \times N$ area of pixels surrounding one pixel in the left image be compared to an $N \times N$ area of

* This is analogous to "local stereopsis" in man (Ref. 22).

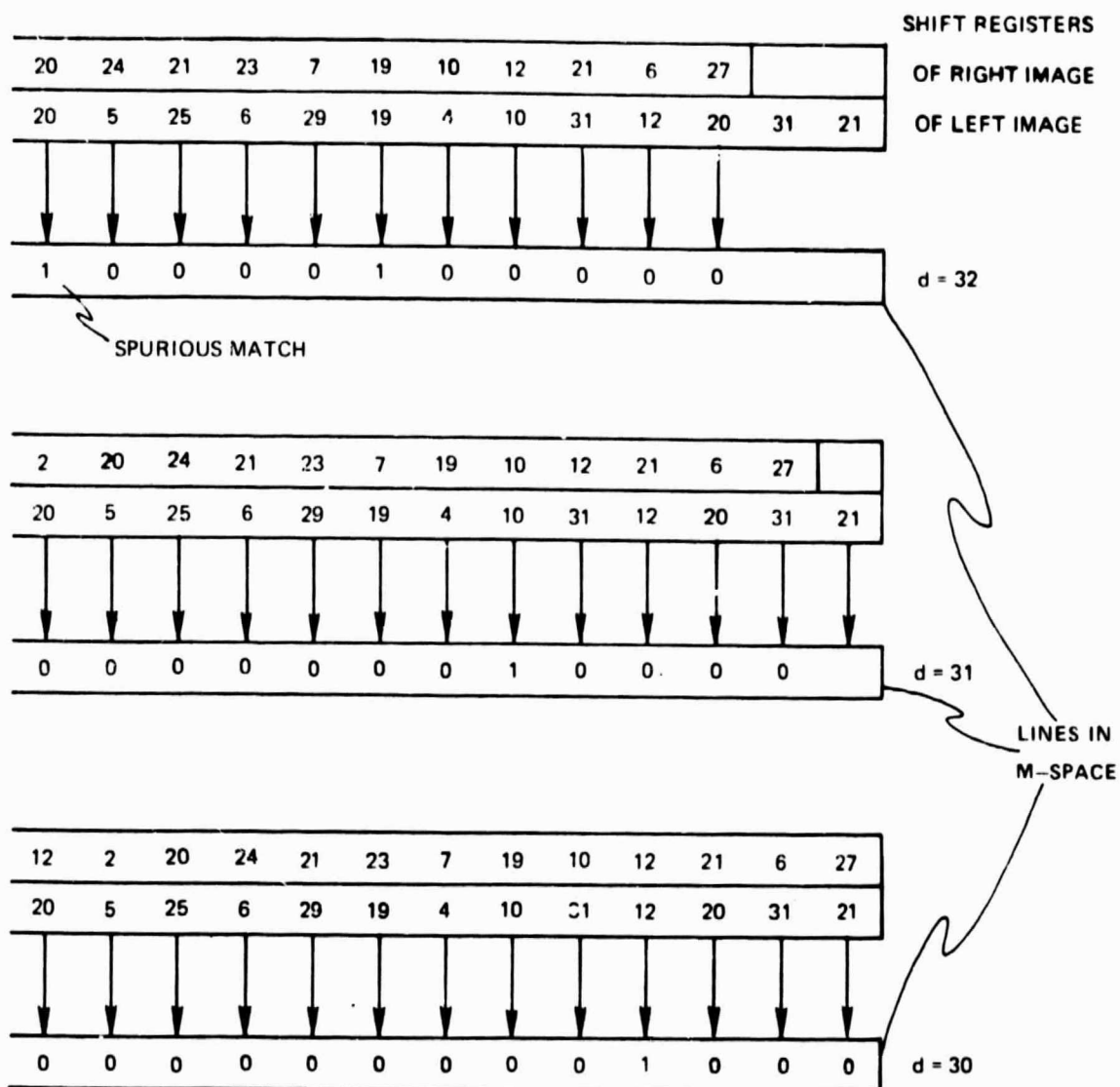


Fig. 5. Computation of first-stage matches at the right ends of three lines of the x' - d plane picture, in Fig. 3. The tolerance of first-stage match, ϵ , is here 0.

pixels surrounding a second pixel in the right image. The number of first stage matches required for a second stage match is $\geq N^2$ (LOWLIM/100), where LOWLIM/100 is a preselected percentage of the first stage matches. Values of N and LOWLIM/100 used in the examples of this paper are shown in Table 1. As will be explained in II K, the program EXPER increases the search area beyond N^2 to resolve "ties", i. e., surfaces of the same number of first-stage matches, in an attempt to find the larger area.

The above description glosses over the step that preceded the simulation of the matching of the left and right images of the Mars-like scene pictured in Fig. 16. First, the left image of Fig. 16 was digitized and stored on magnetic tape. The camera was then moved 50.8mm. to the right and the right image digitized and stored on magnetic tape. A human operator then determined the vertical disparity (y shift) of the left and right images by matching the digitized video waveform of one scan line of the left image with the digitized video waveform of a scan line of the right image with the aid of the program MSTUDY. The scan lines he compared were of a region of uniform range, namely, the flat back drop. The operator eliminated the vertical disparity by entering its amount, usually only a few lines, into the program STEREO.

STEREO forms on magnetic tape the match space required by the next simulation program. (Magnetic tape is used only in the simulation.) The proposed hardware, described in Section III, would hold only as many digitized lines of an image as are required by the filters that examine them. It would store only N planes of m -space.

The need for the y shift is eliminated in the C3 camera described in Section V, but not in the D1 or E1 cameras. For the latter two, either the amount of the vertical disparity will have to be computed automatically and used to match the two images vertically or the next stage of computation must be designed to tolerate vertical

TABLE 1
VALUES OF THE MATCH VARIABLES

<u>Subject</u>	<u>ε</u>	<u>N</u>	<u>LOWLIM/100</u>
	in STEREO program	in STROUT program	
Random-dot stereogram (Fig. 9)	0	3	.50
		in EXPER program	
Random-block stereogram (Fig. 14)	0	3	.50
Mars-like scene (Fig. 16)	4	13	.50

disparity. The nervous system of vertebrate animals tolerates some vertical disparity in computing depth (Ref. 23) and interprets this disparity in the "induced effect" (Ref. 34).

The contents of m -space can be displayed in the form of either x' - d sections, such as those of Figs. 10, 11, 12, 15 and 17, or in the form of a range map. An x' - d section of m -space can be generated for any value of y by the program MSTUDY. The range map is described in II I.

F. Use of Model in Match Space

After a model has been formed in match space at least two questions need to be asked: (1) Is each match plotted there probably true or probably spurious? (2) Are there surfaces in the original scene that are viewed by one camera, not the other and are therefore not modelled in match space? We call such surfaces "ambiguous".

Subsections II H to II L present the rules we have devised to answer the above questions. The rules evolve in three stages and are illustrated by scenes of increasing complexity.

The simplest scene is one generated by computer from dots of random values of gray, briefly called "random dots". (Our usage here differs from that of Bela Julesz (Ref. 24). He uses the term "random dot" to describe a two-value, black-and-white display.) Such a scene is simplest because the probability of a spurious surface in match space is negligible.

A scene of higher order complexity is again computer-generated but now of random uniform areas of gray. We call such scenes "random-square" or "random-block" pictures. Here the probability of a spurious surface of matches is greater, because any spurious event will automatically be a surface.

A third order of complexity is a real scene. This can be processed by the rules devised for random square pictures.

G. Geometry of Binocular and Match Spaces

Before proceeding with rules for interpreting a model in match space we need to take another look at the geometry of a stereo TV camera assembly. Figure 6 diagrams the same two cameras pictured in Fig. 3, the same binocular and match spaces, except that the views are now from behind the cameras. The disparity between the images S_L and S_R of the point P is the difference, $d_L - d_R$, whether the object is between the optical axes, as in Fig. 6 or at one side of them as in Fig. 7.

The example for which range needs to be computed is the Mars-like scene of Fig. 16. The range, z , of any point, P , measured from the optical centers of the lenses, is

$$z = \frac{2bf}{d} \tag{1}$$

where the variables have the meanings given in Fig. 6. (For a derivation, see Appendix A. 1.) There is an uncertainty, Δz , in this measurement for which an equation is derived in Appendix A. 2.

In the system, the operation of which we describe here, the number of pixels in each line scanned by the TV camera for both left and right views is 512. That is, the counter that determines the position of the electron beam along the horizontal axis of each TV image counts to 512. The counter that determines the vertical position also counts to 512. However, only the central 256 columns and 256 scan lines were used in acquiring the images of Fig. 16. Each computer-generated image used as an example in the next three subsections also measures 256 x 256 pixels. Only 128 columns, approximately at the center of these images, were compared and the matches plotted in the x' - d sections shown in Figs. 10, 11, 12, 15 and 17.

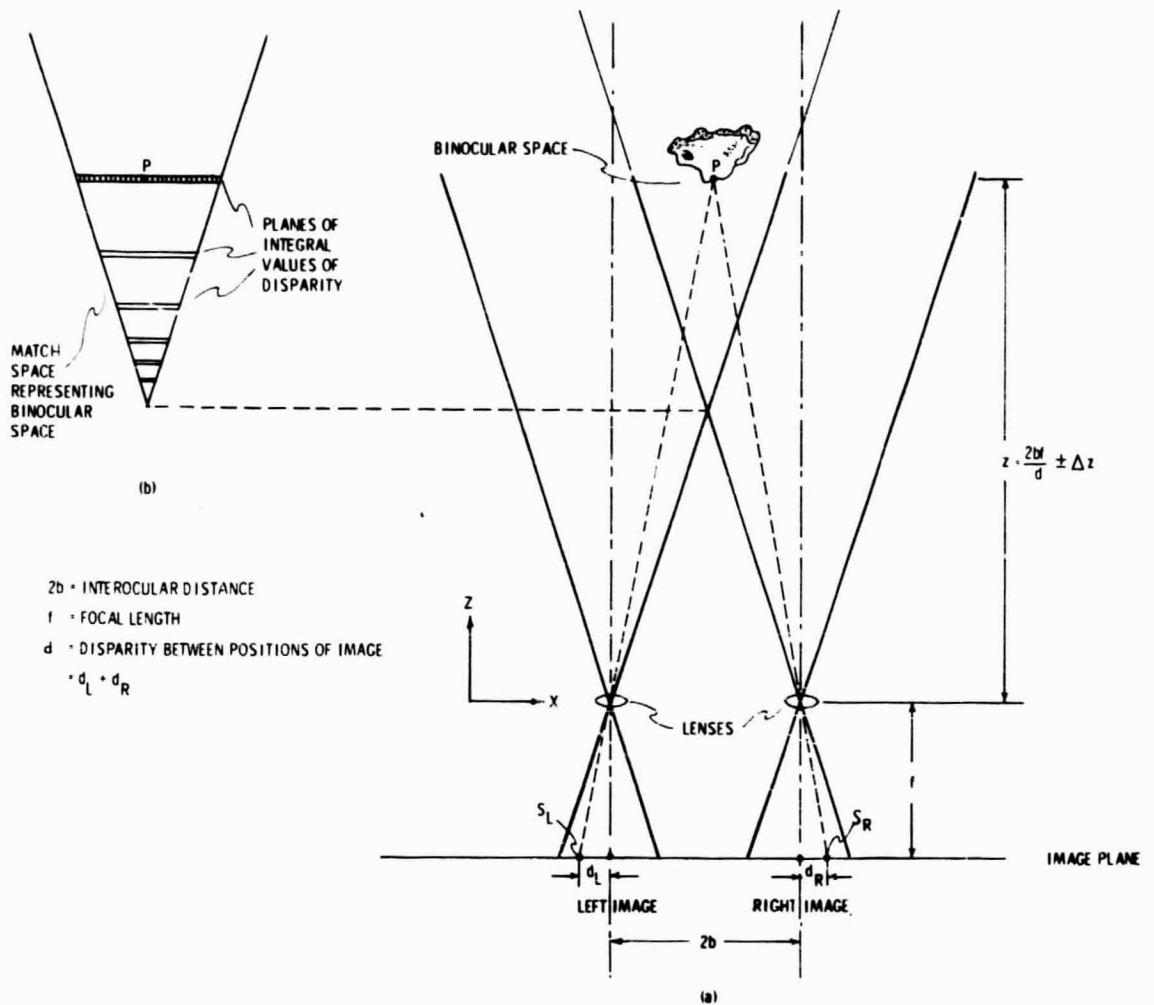


Fig. 6. (a) Geometry of Fig. 3, when object viewed is between parallel optical axes.
 (b) Match space corresponding to the binocular space in (a).

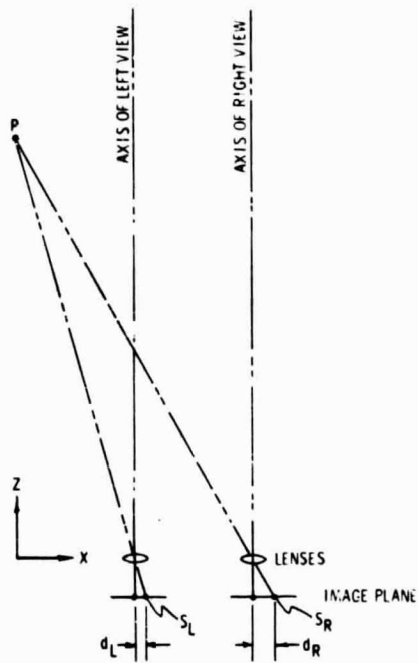


Fig. 7. Geometry of Fig. 3 when the object viewed is at the left of both optic axes.

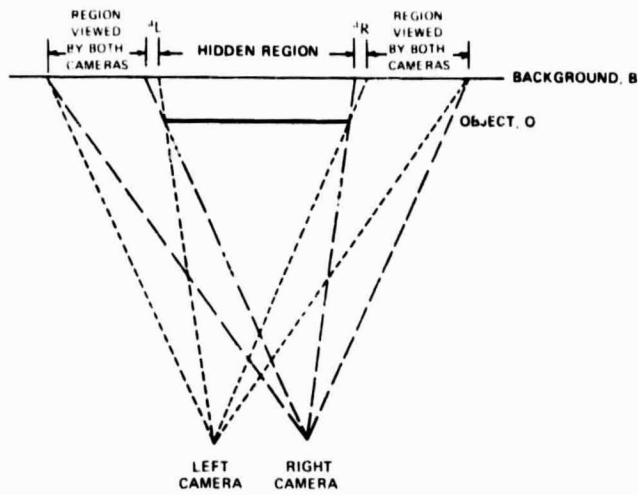
H. Rules for Processing Random-dot Stereograms

Random-dot stereograms were generated in our experiments to investigate the detection of spurious matches and the elimination of ambiguous regions. Figure 8 is a plan view of objects, O, which hang in space before backgrounds, B. Figure 9* is a stereogram of a scene like that in Fig. 8 formed from identical random dot patterns. A 64 x 64 pixel region in the left background was shifted 5 pixels to the right so that it appears nearer than the background. The region uncovered by the shift was filled with more random dots.

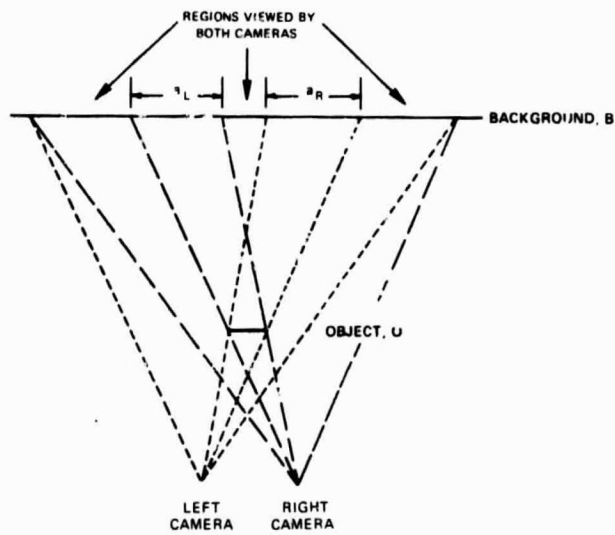
The simulation program STEREO compares the images of Fig. 9, makes first-stage matches and maps them in center-view m-space. STEREO produces one plane of first stage matches at $d=0$ corresponding to the background, a second plane of matches at $a=5$ corresponding to the square floating in space and some spurious matches. Figure 10 shows an $x'-d$ section of this m-space. (The checkered area of m-space in Fig. 3 is a similar section.)

The simulation program STROUT examines each "region" which is a volume extending from front to back of m-space, N pixels high and N pixels wide, seeking a surface of at least $N^2(\text{LOWLIM}/100)$ matches in a plane perpendicular to the optical axes of the cameras. A more complex simulation program (and a more useful one) would search for arbitrarily oriented surfaces in m-space. When, in examining the stereogram of Fig. 9, STROUT finds a plane that meets the conditions $N=3$, $\text{LOWLIM}=50$, it maps the point in a "range map". By this we mean a one-eyed view of the scene in which each

* A good quality viewer for the stereograms in this report is the Model PS-2, made by Air Photo Supply Corp., 158 South Station, Yonkers, New York, 10705. Use 63mm separation of lenses in this viewer unless your eyes are closer together or further apart. Most copies of this report contain a viewer made of cardboard and plastic lenses. This viewer can be purchased only in a large quantity.



- a) Scene pictured in the stereogram of Figure 9 and modelled in the m-space of Figure 10. Region a_L is ambiguous because viewed only by the left camera. Region a_R is ambiguous because viewed only by the right camera.



- b) Scene modelled in the m-space of Figure 11b. The object is now so small and near that the cameras view the background behind it.

Fig. 8. Why the range of the parts of the background, at the left and right sides of an object, is ambiguous.

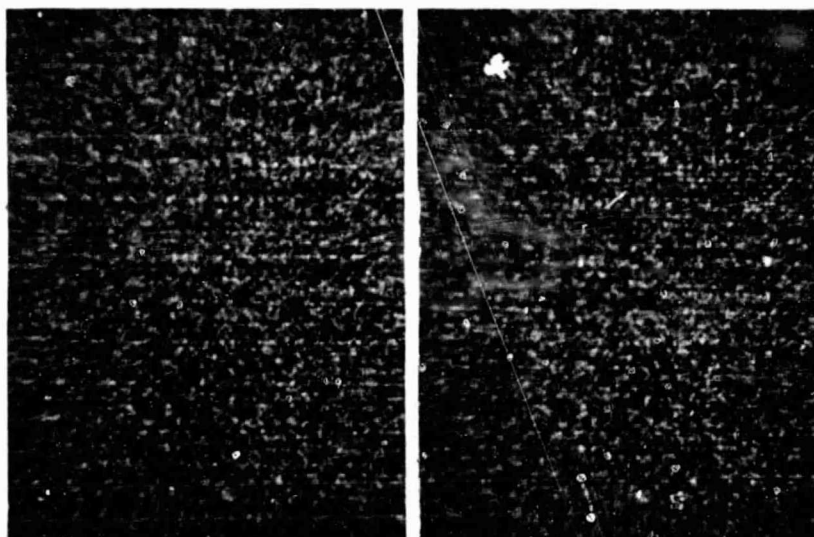


Fig. 9. A random-dot stereogram depicting a square floating before a background.

point represents, by a gray value, the disparity (or range) of the nearest surface in the scene at that pixel.

Because it requires that a percentage of the points in an $N \times N$ plane be first-stage matches, STROUT rejects isolated first-stage matches as spurious. When the image of a plane perpendicular to the z axis (Figs. 8 and 9), is modeled in m -space (Figs. 10 and 11), all the matches lie in one $x'-y'$ plane. Thus a spurious match in a given $x'-y'$ plane will not have many neighboring matches in that plane, while a second stage match will adjoin other matches.

What should be done in a region of m -space where there is no second-stage match is simple for the examples given in this report, but can be complicated for other examples. Let us consider first the simple examples so far presented. STROUT labels as ambiguous regions of m -space where no match is found (Figs. 8, 10 and 11). A subroutine RESOLV then follows the observation of Julesz (Ref. 25) about simple stereograms such as that in Fig. 9: "Regions of ambiguity are always perceived as being the continuation of the adjacent area that seems farthest away". RESOLV searches m -space one plane at a time for regions marked as ambiguous. When it encounters one, it examines the first non-ambiguous region to the left and right and chooses the one which represents a surface farthest from the camera to replace the marked ambiguity. The regions a_L and a_R (Figs. 8 and 10) will then have been filled in and the ambiguity removed.*

This method of treating ambiguous regions is appropriate when the object viewed is a plane, as in Fig. 9, but suppose the

* This process is analogous to what Julesz calls "global stereopsis". To quote him: "With increased dot density the visual system cannot find uniquely the corresponding points, and a new process has to be invoked which can resolve ambiguities by global considerations." (Ref. 22)

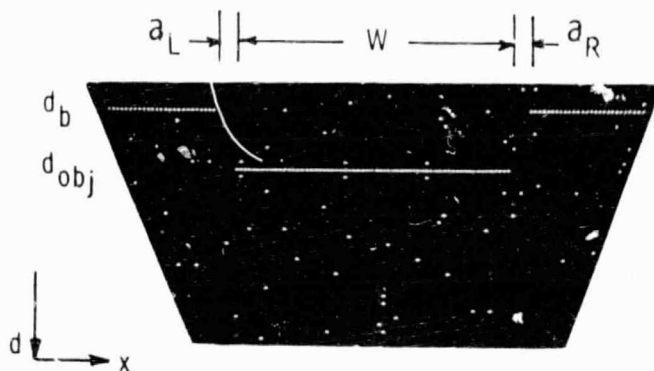


Fig. 10. An x' - d section of the m -space generated from the stereogram in Fig. 9.

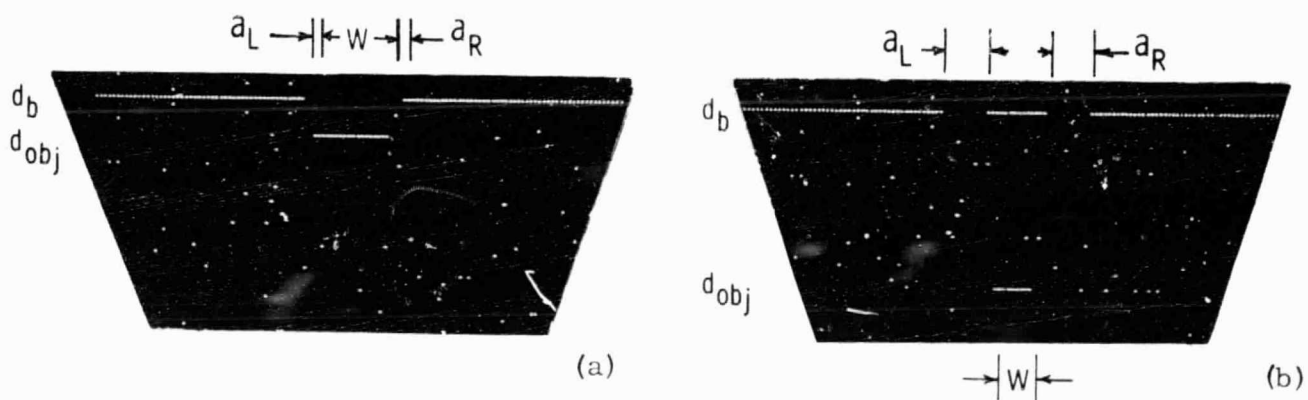


Fig. 11. x' - d sections of m -space showing, in (a), ambiguous regions a_L and a_R on the left and right sides, respectively, of the model of the object and, in (b), how both the image of the object and the background behind it are modelled when $w < (d_{obj} - d_b)$.

object is solid and the left side is viewed only by the left camera and the right side only by the right? For such a scene, another technique is needed; namely, one in which the memory of textures and patterns, viewed before, is applied to interpreting what can be viewed by only one camera. This is cognitive processing.

I. How Should Ranges in a Scene Be Presented?

In building a sequence of computations into a process, such as the detection of range and shape, means are needed of viewing the steps in the process. Such a means is a projection onto an $x'-y'$ plane of all values of disparity in m -space. Since an array of numbers is difficult for a human being to interpret, we have written a program PICT which displays these numbers as levels of gray on a cathode ray tube where they can be photographed. Figure 18 is such a photograph. While the levels of gray correspond to measurements of disparity, the impression given the viewer is that of range. Hence we call this display a "range map".

After using range maps for several years we find that they suffer a defect. When, as in Fig. 8b, the object O is small and near enough to the cameras that they view both the object and its background, the range map is unable to report both. In Fig. 8b the region between the two ambiguous regions, a_L and a_R , is visible to both cameras, but the range map cannot show it.

A stereogram range map could obviate this difficulty. However, if EXPER did not eliminate the more distant of two surfaces in the same region, it might not eliminate spurious area matches either. Therefore, for the present, we accept the limitation that a surface behind a front surface cannot be shown in a range map.

J. Eliminating Areas of Spurious Matches

Because areas of spurious dot matches are not likely to be

formed from a random-dot stereogram, STROUT is not equipped to reject such areas. Such areas are likely to be formed both from random-square stereograms (Fig. 14) and from stereograms of real scenes (Fig. 16). The latter tends to contain areas of the same value of gray (within tolerance ϵ) because its dots are not random.

To eliminate areas of spurious dot matches we devised two simulation programs, EXPER and FUSER. Let us consider with the aid of Fig. 12 how EXPER might be used alone for this task. It will be shown in Fig. 15 that a model in m -space of an area of uniform gray in a scene appears as a parallelogram in an x' - d section of m -space. Because parallelograms are awkward to illustrate in this stage of our discussion, we will continue to use a line to represent a plane in Figs. 12 and 13.

Assuming again that the background of the scene pictured in a stereogram will be continuous, we can see how right-view and left-view m -spaces can be used to reveal whether or not a surface is spurious. Consider again the scene mapped in Fig. 8b. Three forms of the simulation program STEREO form, from a stereogram of this scene, left-view, center-view and right-view models in m -spaces. (The program MSTUDY generates the x' - d sections of these m -spaces shown in Figs. 12a, 12b and 12c.) Because, in Fig. 12a, the model of the object hides the right ambiguous region, and, in Fig. 12c, the model of the object hides the left ambiguous region, the model of the object is considered true, not spurious. To make this check automatically, the data of the left-camera and right-camera views needs to be converted to center-view m -spaces and these m -spaces compared. That will be done by the program FUSER.

EXPER performs more operations than we have so far described, as Fig. 13 shows. Figure 13 begins at (a) with the plan view of a simple scene: an object O in front of a background B.

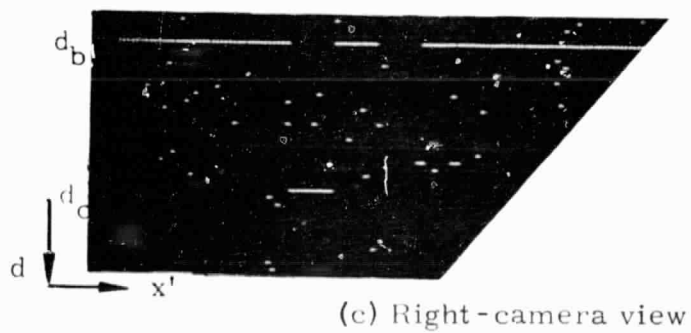
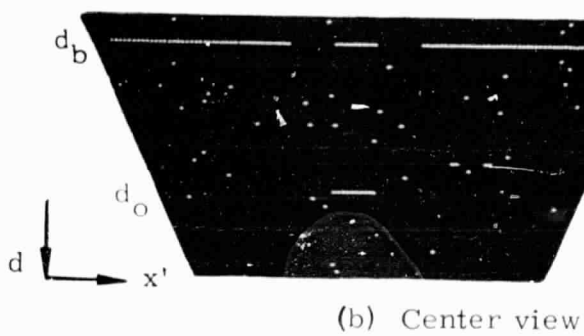
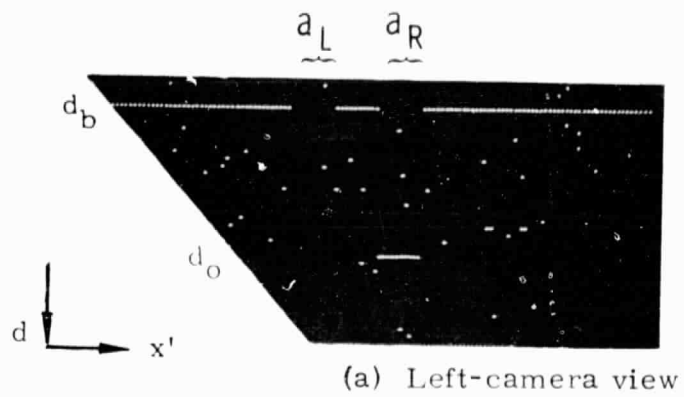


Fig. 12. x' - d sections of three match spaces formed from the same stereogram.

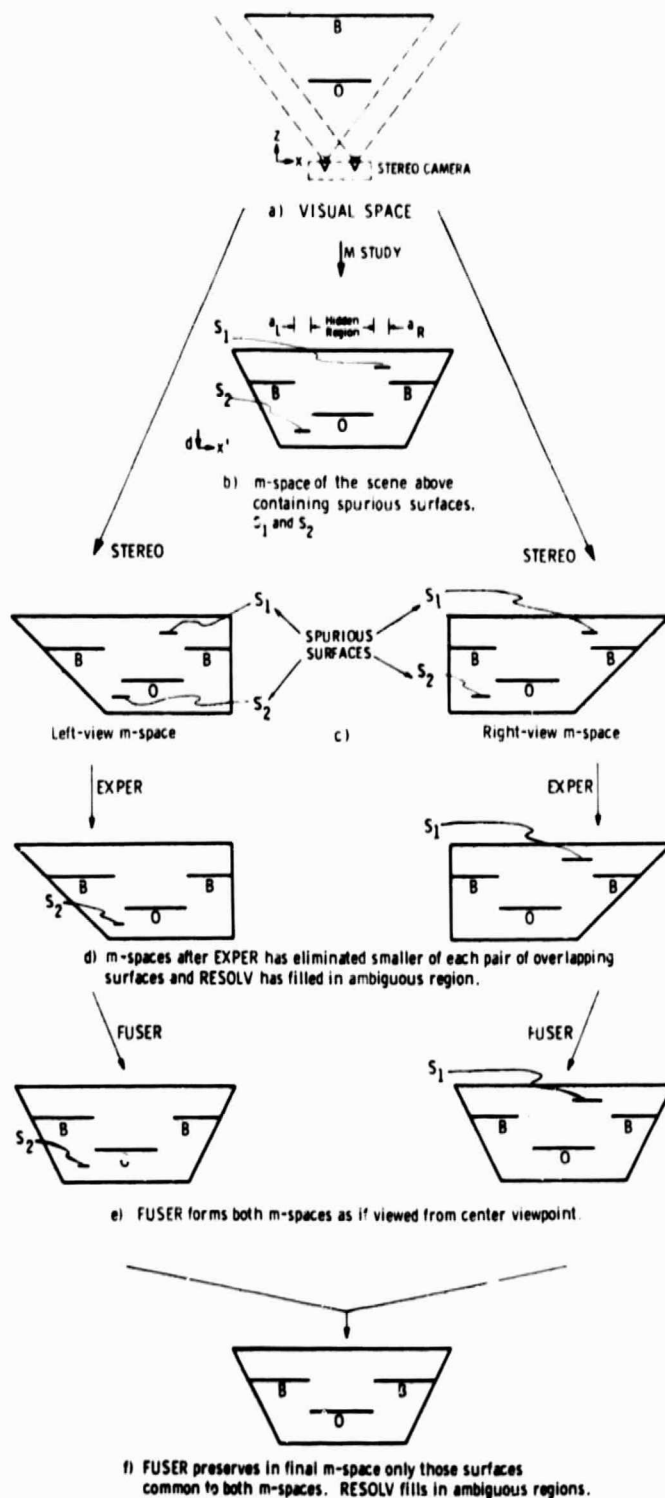


Fig. 13. Diagram showing the functions of the three simulation programs, STEREO, EXPER and FUSER.

Step (b) is performed by MSTUDY to show an x' - d section of center-view m -space and two spurious matches, S_1 and S_2 , which happen to hide the two ambiguous regions a_L and a_R . STEREO forms left-view and right-view m -spaces of which MSTUDY forms the x' - d sections shown at (c).

Where there are two or more surfaces in a region of either left-view or right-view m -space, EXPER eliminates all except the largest in the following manner. At each set of values of x' and d , EXPER begins by searching an $N \times N$ region, in this case 3×3 , for possible surfaces. If two or more are encountered, EXPER enlarges its area of search until it finds the largest. It then retains only the point on this largest surface. For example, in the left-view of m -space (Fig. 13c), because S_1 is smaller than O, EXPER eliminates S_1 ; and, because S_2 is smaller than O, EXPER preserves only those points in S_2 which do not hide points in O. In the right view of m -space, because S_1 is smaller than B, EXPER preserves only those points of S_2 not hidden by B; and, because S_2 is smaller than B, EXPER eliminates S_2 .

Employing the subroutine RESOLV, described in II H, EXPER, in the right-view m -space, extends S_1 leftward into the ambiguous region between O and S_1 . The results of these operations by EXPER are shown in Fig. 13d. Note that S_2 survives in the left-view m -space and S_1 has grown in the right-view m -space. Thus EXPER alone cannot eliminate all areas of spurious matches.

A simulation program that compares the EXPER-processed left- and right-view m -spaces to complete the elimination of areas of spurious matches is FUSER. From each m -space illustrated in Fig. 13d, FUSER forms the two center-view m -spaces shown at (e). FUSER then compares these two m -spaces, preserving only those surfaces common to them both. Finally, FUSER employs RESOLV to fill in remaining ambiguous regions.

The assumption is that each surface in the scene is opaque, and therefore blocks out, behind it, one region from the left camera and another region from the right camera. Forming left-view and right-view m-spaces is an attempt to check for this condition. Comparison of what is mapped in the left-view m-space with what is mapped in the right eliminates surfaces that do not satisfy this constraint.

K. Random-Square Stereograms

Figure 14 is a random-square stereogram of a large square floating in front of a background. It is formed of random-gray-value squares measuring 4 pixels on a side. Figure 15 is an x' -d plane of center-view m-space formed by MSTUDY from this stereogram.

The match of left and right views of each random square of Fig. 14 is a diamond because each square in binocular space is perpendicular to the z' -axis. If the square is skewed with respect to this axis, the match is a parallelogram. The shape, whether a diamond or parallelogram, is formed as the left view marches past the right in the scheme of Fig. 3. At the first overlap of left and right images of a uniform area of gray, the point of the parallelogram is formed. After a shift of the right image, the next line of the parallelogram is formed, two pixels wide, at a larger value of disparity. When the two small squares overlay each other, the widest part of the parallelogram of matches is formed.

In its search for the largest number of matches along each y' -d line EXPER finds the widest part of each diamond and retains only that. Thus the diamonds of Fig. 15 are reduced to the lines of Fig. 13.

L. Processing of a Real Scene

Figure 16 is a stereogram of the scene of Fig. 1 recorded when

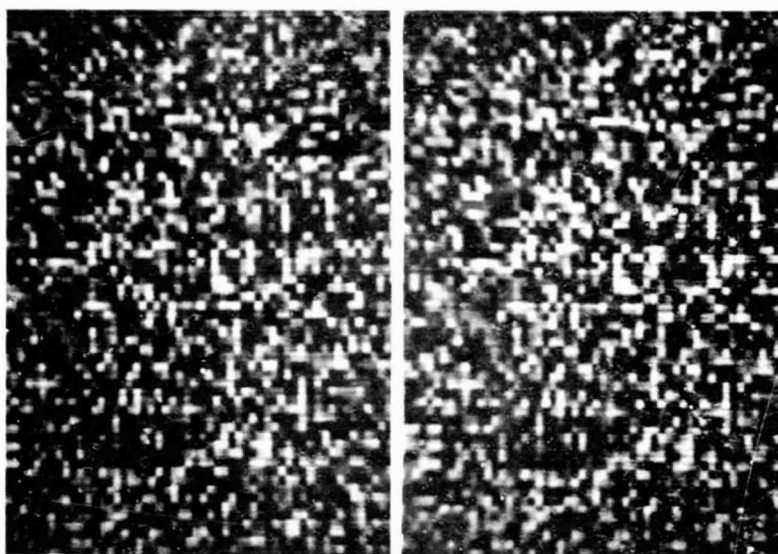


Fig. 14. Random-block stereogram of a square floating before a background. Each block is 4×4 pixels of uniform gray value.

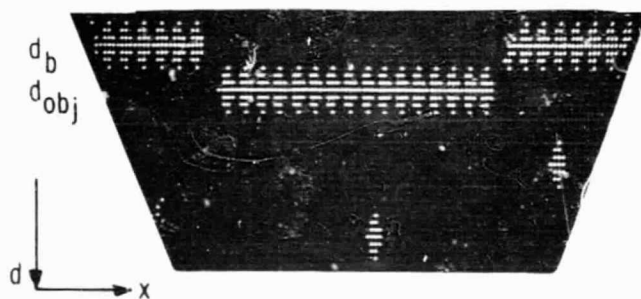


Fig. 15. x' - d section of model in m -space generated from the stereogram in Fig. 14.

the TV camera was approximately 2m from the rock and the lighting was from upper right. Other test conditions are given in Tables 1 and 2.

One camera was used to obtain the left view, then moved 50.8mm (2 in.) to the right to obtain the right view. The axes of the camera in its two positions were parallel. Since the axes of the cameras were not changed, this was the fixation viewing of the second strategy of II B. Since a lens of 25.4mm focal length was used, the resolution was that of the first strategy. Such a compromise is necessary when only one focal length is employed.

Figure 17 is an x' - d section through the left-justified m -space formed by STEREO from 128 columns of a stereogram similar to Fig. 16. The section is at the y -value of the stereogram indicated by the two black lines in the margins of Fig. 16. One of the diamonds in Fig. 17 models the stick. Another may model the shadow of the stick. The right side of Fig. 17 contains at the front spurious matches and further back true matches of features in the painted backdrop.

Figure 18 is a range map formed by EXPER from the m -space of which Fig. 17 is a section. Lightness of gray indicates nearness, darkness of gray, distance. Thus the rock stands out clearly. Two shades of gray at the top of the rock, one shade hooking to the right, indicate how long the rock is. The stick and its shadow get progressively darker as it recedes, leading to the backdrop which is a uniform black. Because occluded regions have not yet been found and filled in, two of them appear as black areas below the rock and below the stick. A picture was made of the output of FUSER after it had filled in these occlusions, but it was poorly displayed so we do not include it.

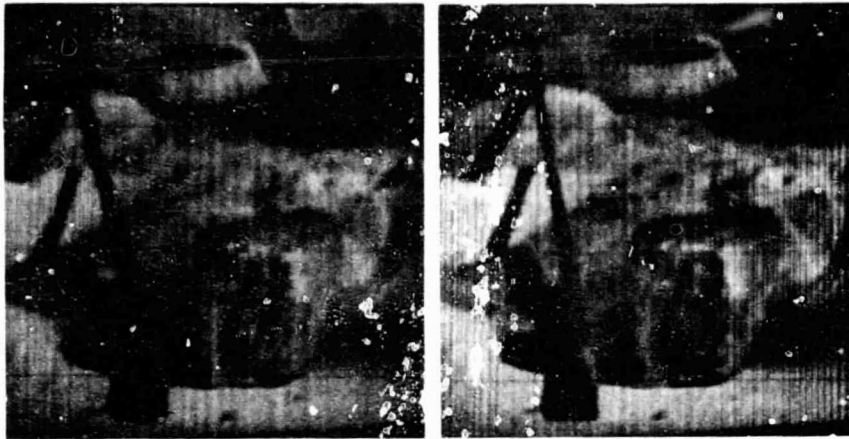


Fig. 16. Stereo images of the Mars-like scene of Fig. 1, acquired by television camera, digitized, then displayed one at a time on an oscilloscope.



Fig. 17. x' -d section of the right-camera view of match space formed from the stereogram of Fig. 16. The y value of this section is indicated by lines in the margins of Fig. 16

TABLE 2

TEST CONDITIONS FOR EXAMPLE OF H L

(In addition to those given in the bottom line of Table 1)

Interocular distance, 2b	= 50.8mm (2 in.)
Focal length, f	= 25.4mm (1 in.)
Distance from camera to nearest object	= 2m (approx.)
Camera tube type	= GEC TD8484
TV camera and control	= Colorado Video, Inc., Type 501
Computer	= Digital Equipment Corp. PDP-9 with 8K words of core memory and 2 DEC tape drives
Display oscilloscope	= Tektronix 530
Display-tube type phosphor	= P4

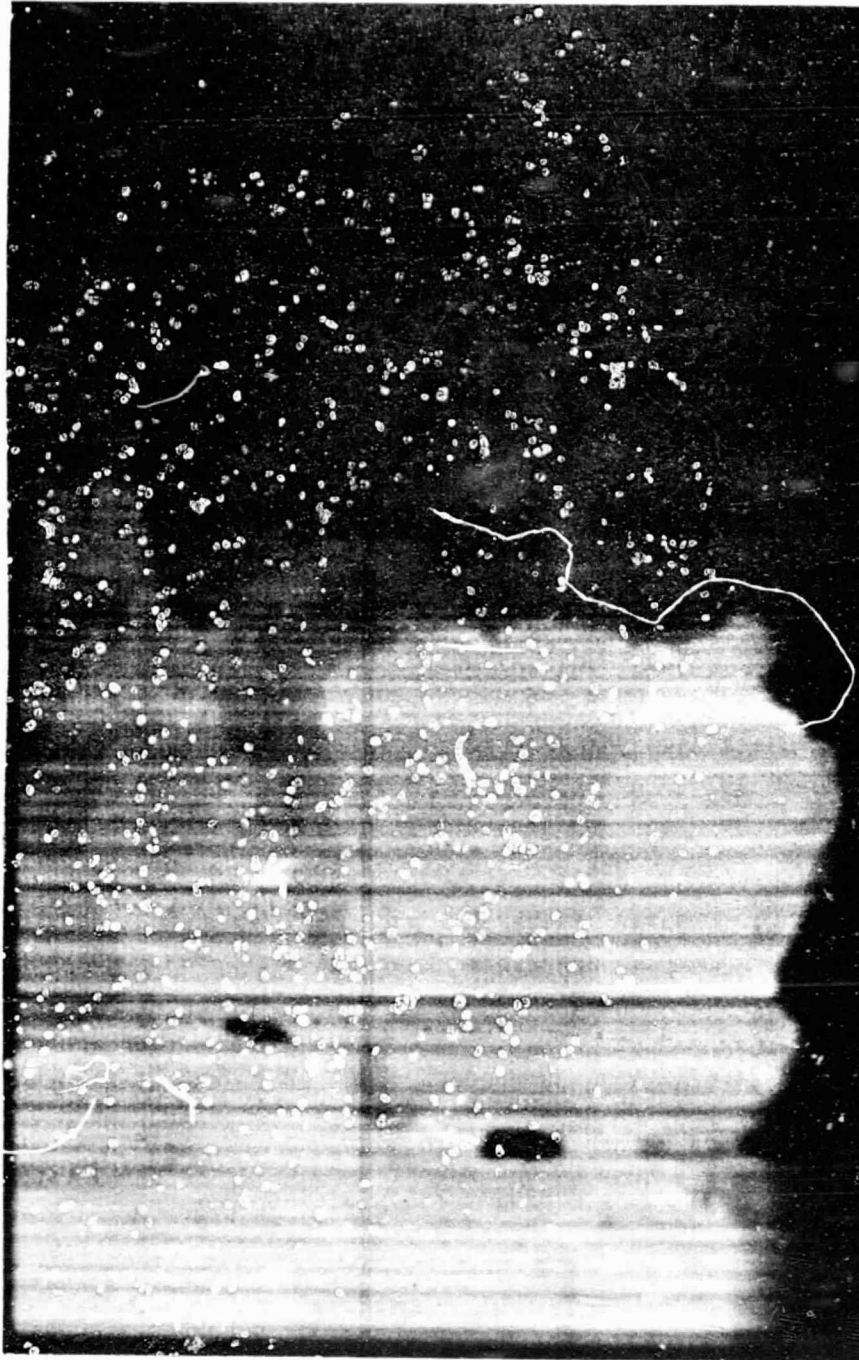


Fig. 18. Range map of one match space of the stereogram in Fig. 16. The black areas below the rock are ambiguous because they were viewed by the camera in one position and not in the other.

M. Determining Form

McCulloch observed that texture and form are not primarily visual phenomena. They are "the way a surface or object would feel if you could feel it" (Ref. 26). To determine form this way range data should be fed both to a touch system and to the controls of an arm and hand that is capable of reaching into the space viewed by the stereo TV camera. Since judgement of form from camera input data will be only a guess, the arm and hand can confirm or deny this guess. Design of a system to operate this way is considered in the final paragraph of V.

N. Range Accuracy

How accurate is the above system, assuming that the ambiguities just described have been removed and the spurious matches eliminated? How can range accuracy be increased?

The stick in Figs. 1, 16 and 18 is 1.27m (50 in.) long. In the original Polaroid print of Fig. 18, there are seven levels of gray along the length of the stick. Dividing seven into 1.27 indicates that intervals of range have been detected of about 18cm (7.1 in.). This is approximately the uncertainty that is predicted when measured characteristics of the camera are substituted into Eq. (2) below. The rock, the length of which is 20cm from front to back, is shown as two levels of gray.

Appendix A derives the following formula for range uncertainty

$$\Delta z = \frac{\Delta s \cdot z^2}{bf - \Delta s \cdot z} \quad (2)$$

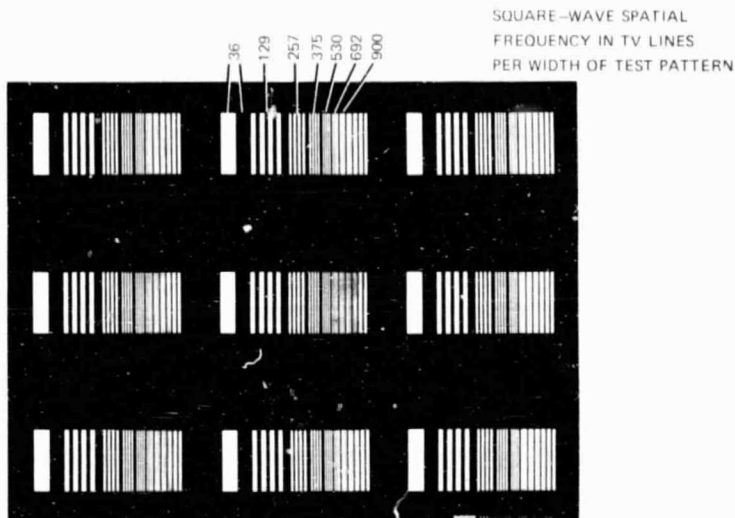
Because Δs is very small with respect to z , the second term of the

denominator may be ignored. Eq. (2) then becomes

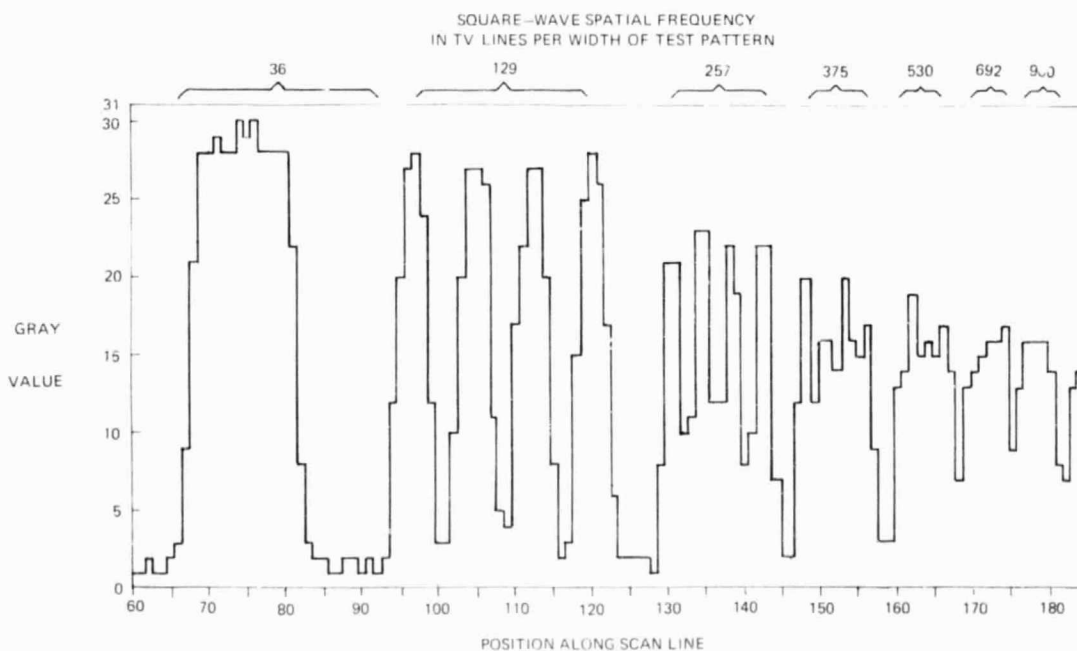
$$\Delta z = \frac{\Delta s \cdot z^2}{bf} \quad (3)$$

Assuming that the camera tube is selected for the smallest possible uncertainty, Δs , in the position of a point, it can be seen that range accuracy can be increased either by increasing the focal length, f , of lenses, by increasing the separation of the optic axes, $2b$, or by bringing the camera nearer to the objects to be examined (smaller z). If longer focal length lenses are used to examine details, short focal length lenses are still needed as finders of the details to be examined (strategy 1). Thus, a stereo TV camera is needed with lenses of two focal lengths. If wider separation is used between the optic axes, the axes need to be converged on the objects of Fig. 16, requiring trigonometric functions in Eq. (2). Such functions can be employed but were avoided in this first pass through the problem. Obviously the camera could have been brought nearer, but then it could not have viewed as many objects as in Fig. 16.

Increasing by a factor of ten the focal length of the camera lens employed in the test described in II L will reduce range uncertainty to 1.8cm (0.7 in.). Increasing the interocular distance by a factor of 4, to widths described in V, will further reduce the range uncertainty to 0.45cm (0.18 in.). However, these changes can bring other problems. Lenses of focal length this long cannot be accommodated in the Type C3b camera. While such lenses can be accommodated in the Types D1 and E1, there is a problem of vertical misregistration (vertical disparity) in these types, as explained in II E, which special computation is required to remove. Solutions to these problems are being devised.



a) "Line Selector" test pattern (Westinghouse resolution chart ET-1332 purchased from Tele-Measurements Inc., 145 Main Avenue, Clifton, N. J.), reproduced 1/3 full size.



b) Plot of digital words formed by the system of Fig. 0 as the electron beam in the camera sweeps the image of (a).

Fig. 19. Test pattern and amplitude response of TV camera

III. COMPUTATION TO EXTRACT OTHER FEATURES

A1. Square-Wave Frequency Response of Camera

Other features, besides range and form, that may be detected in a scene, are the edges and lines shown to be detected by cats and monkeys (and probably also by human beings), and the reflecting properties of surfaces of interest.

The computation of an edge by a TV camera-computer system needs to be described in the language of TV cameras, computers and picture processing. A TV camera makes a "square-wave response", which is converted into "digital words" for the computer. Each word indicates the "gray value" of a "pixel". An edge is a difference in gray values (Ref. 27).

To determine the square-wave frequency response of the camera we replaced the stereo optics of Fig. 0 by a single lens and aimed the camera at the transparency of Fig. 19a which we lighted from behind. We positioned the camera so that the transparency and its margin were just included within the scanned area of the camera tube.

Figure 19a provides spatial square waves of 11 different frequencies, seven of which are labelled. A square-wave spatial frequency, to a TV camera, is the number of lines, both "black" and "white", that can be imaged in the scanned area of the camera tube. We measure this frequency in TV lines per width of test pattern. Because there are both a "black" and a "white" TV line in each cycle, the number of cycles per width of test pattern is one half this number.

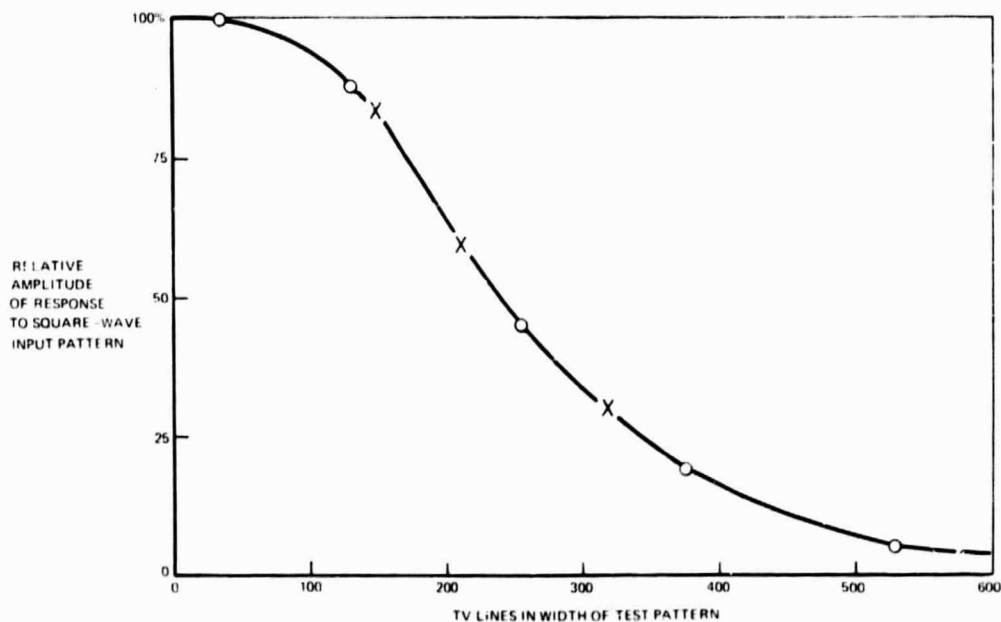
As the electron beam in the camera tube scans along one scan line, the camera generates a voltage which an analog-to-digital converter changes to digital words. The voltage from one scan line is converted to 512 digital words, one for each of 512 positions along the scan line.

Figure 19b plots digital words formed as the beam sweeps through 120 of these positions. (Of the 512 positions in the camera's image of Fig. 19a, the central 256 are presently acquired by the computer. Of these, positions 60 to 180 are plotted in Fig. 19b.)

Each step in Fig. 19b is a pixel whose gray value is indicated by a digital word on a scale from 0 to 31. Figure 19c plots the average change in amplitude between the black and white halves of each square wave. The plot is made relative to the 36 TV lines/frame frequency. From this plot can be read numbers that characterize the system, namely, the square wave response, in TV lines, at 10%, 50% and 100% modulation: 470, 230 and 36.

A2. Computation of Edges

Figure 20 pictures the algorithm we designed to detect "coarse" edges. It is composed of six arrays of 0, 1's and -1's which we call



c) Amplitude response. O's are a plot of the relative amplitude of response in (b) to the image of (a). X's mark the square-wave spatial frequencies detected by the top three filters of Fig. 20.

Fig. 19 (Cont'd) Test pattern and amplitude response of TV camera

Number of TV lines that can be detected in image of Fig. 19a:

308

205

146

$$\begin{aligned}
 & \text{Filter} \\
 & \text{Convolution-type process} \\
 & \text{Multiplication} \\
 A = & \left\{ \left[\begin{array}{ccc} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{array} \right] \star \left[\text{IMAGE} \right] \right\} \times \left\{ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & -1 \end{array} \right] \star \left[\text{IMAGE} \right] \right\} \times \left\{ \left[\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 \end{array} \right] \star \left[\text{IMAGE} \right] \right\} \\
 B = & \left\{ \left[\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{array} \right] \star \left[\text{IMAGE} \right] \right\} \times \left\{ \left[\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{array} \right] \star \left[\text{IMAGE} \right] \right\} \times \left\{ \left[\begin{array}{ccc} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{array} \right] \star \left[\text{IMAGE} \right] \right\}
 \end{aligned}$$

$K (A + B)$ = Output for each position in the image

K = Constant

Fig. 20. Operations performed on each 7 x 7 pixel array of a digitized image to detect edges.

"filters". We assign a special meaning to ★, namely, that each number in the filter will be multiplied by the gray value in a submatrix of the image, that is the same size as the filter, and the products summed. (Actually, all filters are made the same size by filling them out with zeros to measure 7 x 7 digits.) To each 7 x 7 submatrix of the image all six filters are applied.

Since our original goal was to form a line drawing from spatial frequency information and our thinking was influenced by the methods of Fourier analysis, we attempted to add each of the terms of Fig. 20. The results were of little value. Then we were advised by Dr. Azriel Rosenfeld to multiply the terms as he did in Ref. 28. The results are shown first in Fig. 21* and, after thinning by searching for local maxima, in Fig. 22.* Multiplication here, as he said, is "counterintuitive," but it serves to detect an edge.

The filter at the upper left of Fig. 20 detects two TV lines (one dark, one light) in three pixels. Across the full width of the pattern of Fig. 19a,

$$2/3 \times 512 = 308 \text{ TV lines}$$

can be detected by this filter, as indicated at the top of Fig. 20. The upper center filter detects two TV lines in five pixels, the upper right filter two TV lines in seven pixels. Along the lower row of Fig. 20 are detectors of the same square-wave frequencies turned 90°. At the top of Fig. 20 are the square-wave spatial frequencies detected by the filters below them. Plotting these frequencies as X's on the graph of Fig. 19c shows the relative response of the camera

* To get the effect of three dimensions, look first at the rock in the foreground, then at the crater in the distance, then at the rock, then at the crater, and so on. Occasionally alter the route by following the stick or exploring the hills.

tube to the frequencies detected.

When all six filters are overlayed, the 0 at their center is the address of the edge that they detect.

B. Formation of A Line Drawing

Plotting all of the edges detected by the algorithm of Fig. 20, in the images of Fig. 16, results in images which are printed in negative form in Fig. 21. That is, sharpness of edge results in brightness of display which is represented as blackness in Fig. 21. While the span of gray values that can be displayed is only 0 to 31, the result of the operations of Fig. 20 is often greater than that. For the display, each result is truncated at 31 so that, for every edge in Fig. 16, there are usually several lines of dots in Fig. 21. We call this a "coarse line drawing."

A "drawing" was thought desirable, when this work began, as a means of transmitting to earth the appearance of a Mars scene with minimum power. The power saving results from the use of a binary code in a raster that is always the same size. The position of a bit in the raster is thus given by the time of its arrival after an initial synchronizing pulse. Figure 22 requires about 1/30 as much power to transmit as Fig. 16.

Figure 23 shows how a coarse line drawing is thinned. Fig. 23a plots the average of the gray values along seven adjacent scan lines in a digitized TV image. There is one gradual transition from light to dark at x_1 , one abrupt transition at x_2 . Both are edges. Figure 23b shows the result of performing the computation of Fig. 20 on the scan line of Fig. 23a. The thinning routine detects local maxima in gray value (A and B in Fig. 23b), locates x_1 and x_2 and displays them on the oscilloscope as the lightest gray (Fig. 22c). We prefer to present the



Fig. 21. Negative of a display of the result of performing the operations of Fig. 20 on the images of Fig. 16.

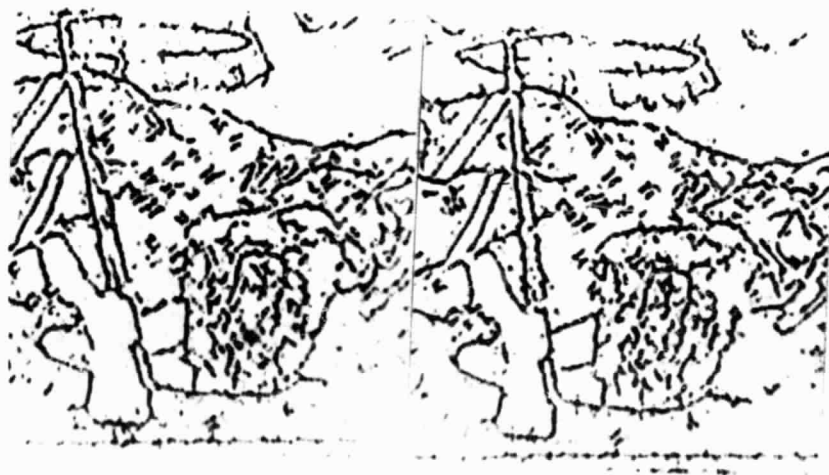
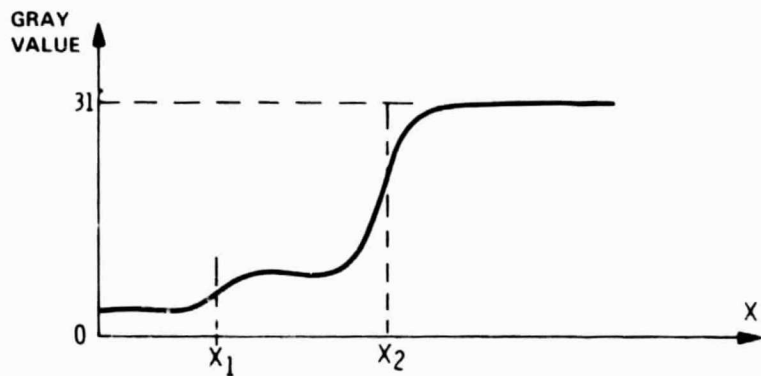
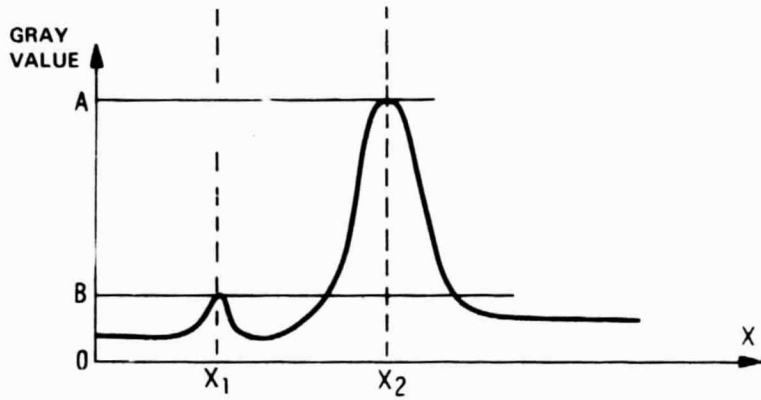


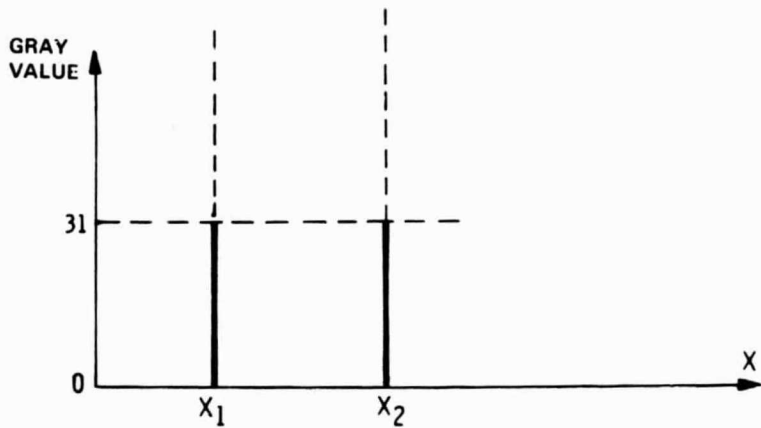
Fig. 22. Edges of Fig. 21 after thinning.



a) Average gray values along seven adjacent scan lines of figure 16



b) Result of processing above lines by method of figure 20



c) Result of detecting local maxima at X_1 and X_2

Fig. 23. Detection of edge and thinning of edge.

negative of such a display. Figure 22 is a negative of the result of applying this edge-thinning operation to the data of Fig. 21.

C. Hardware to Detect Edges

Detection of edges in left and right images can be performed by the assembly shown in Fig. 24. Between the analog-to-digital converters and the comparator pictured in Fig. 3, are two banks of shift registers and computing elements behind each bank. As the electron beam, say, of the left camera tube, detects the signal at the first pixel of a scan line, that signal is converted to a five bit word and fed into the top level of the lower bank of shift registers. As the electron beam advances to the second pixel, the first digital word advances one position along the top level of the bank of shift registers and another word takes its place.

When the electron beam reaches the end of the first scan line, it snaps back to start a second line and the first word that entered the top level of the bank of shift registers is shifted through connections not shown to become the first word in the second level. At the same time the first signal from the second scan line is digitized into a five-bit word and fed into the top level. (Words are shown six bits long because we aim to digitize to this number.) The process just described continues until, when seven levels are full, computation occurs on each 7 pixel x 7 pixel array that passes before the detector of edge (Fig. 24 top center). From then on, after each shift, another 7 x 7 array is processed until the entire digitized image has been processed this way.

Both the gradient of each edge and its polarity (light-to-dark or dark-to-light) were lost by multiplying convolution sums together and taking the absolute value of the product. If, instead, each convolution sum is retained, it can be coded and fed to the comparator in the background of Fig. 24. Comparison between left and right views will then be between gradient, polarity and direction of

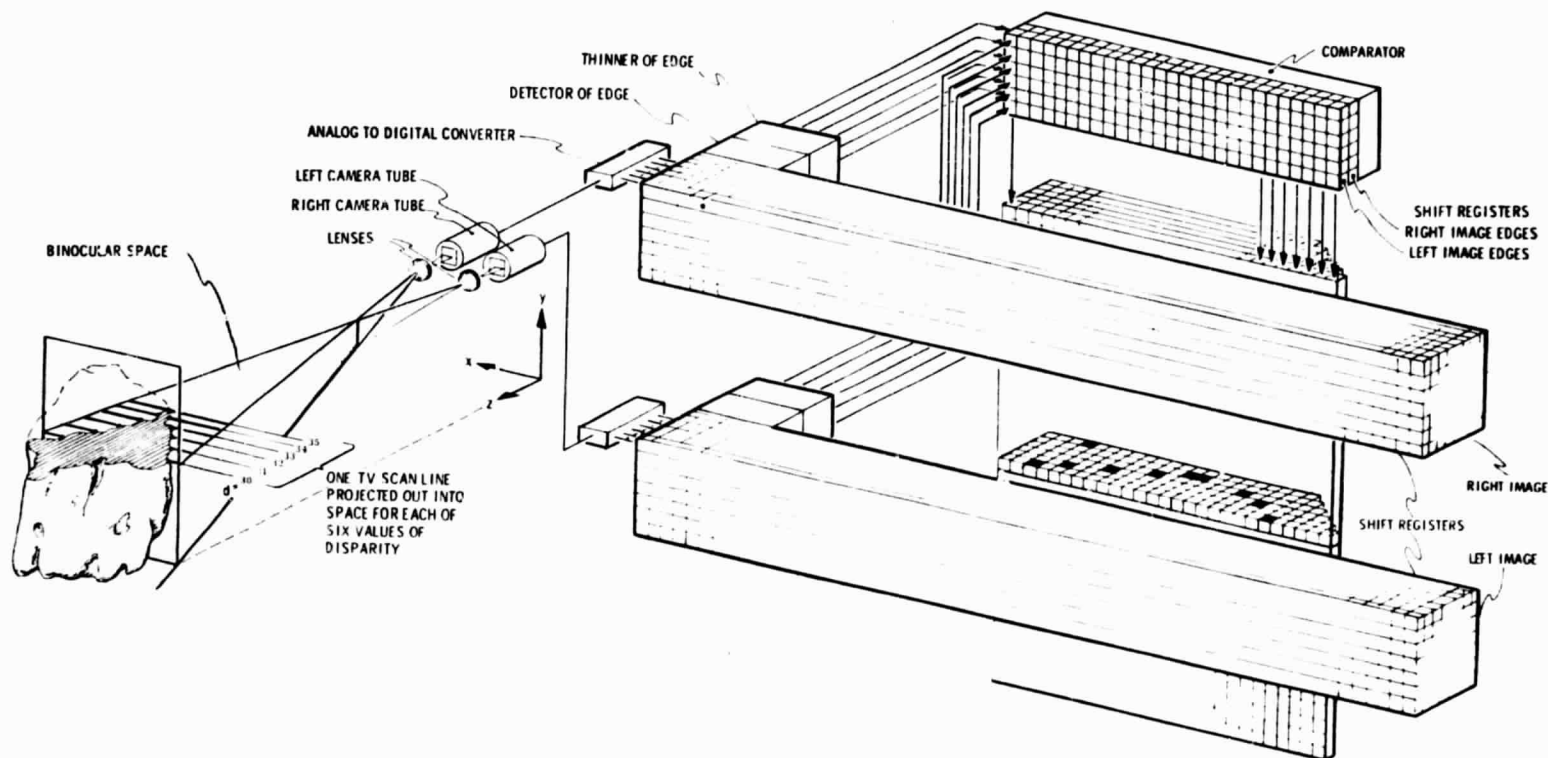


Fig. 24. Insertion of means of detecting edges between the stereo TV camera assembly and match space.

detected edges. When range is computed, it will be the range of a fairly specific edge (or a line, a corner, or other feature). In the opinion of the first author, there should be fewer problems of spurious matches.

Color differences can be determined by employing a three color camera in place of the monochromatic one pictured in Fig. 24 and feeding three digitized color signals into the top-level of each shift register. Computation will then be to detect color differences as well as gray-level differences.

The advantages of the shift registers pictured in Fig. 24 are that they are simple and fast. Shift registers of this kind were designed and are partly constructed (Ref. 29). Performing like cells in the retina and cortex of vertebrate animals, this single detector and thinner of edge is time-shared with the entire area of an image.

A further advantage of separating the effect of each spatial frequency is that each can then be employed in an automatic focussing routine. That is, changes in the low spatial frequency response can be used to indicate in which direction focus can be improved, while the high spatial frequency response can be used to indicate that focus has been achieved.

D. Computation of Reflecting Properties

The reflecting properties of a surface can be determined from the incident illuminance onto a surface and the luminance of that surface. Illuminance, if sunlight, can be measured either by the camera-computer with the face of the camera tube protected by a neutral density filter, or it can be measured by a sun sensor. Luminance can also be measured by the camera-computer. These measurements would be performed through other channels than those pictured in Fig. 24.

IV. RECONSTRUCTING THE APPEARANCE OF A SCENE

By detecting not only features on the surfaces of objects, but also properties of the scene such as the amount and direction of the illuminance, the appearance of a scene on Mars may be reconstructed on earth (Ref. 30). Figure 25 diagrams rays of light from a single source of illuminance, such as sunlight, onto a cylinder. Fig. 26 shows how a computer, using information on the shape of an object, on the sources of illuminance (mainly from upper right, but also from upper left) and on the reflectance can recreate the appearance of that object. The reflectance of the cylinder is here assumed to be diffuse and 100 per cent. The reflectance of the background is assumed to be diffuse and 50 per cent.

Detection of the reflecting properties of a surface, like the recognition of objects, requires more than the passive examination of a scene. There needs to be an active effort to relate incident illuminance to reflected luminance. This is particularly true in the detection of specular reflectance where highlights have to be found and measured. Thus, determination of reflectance needs to be directed by a higher authority than the passive visual computers described in this paper.

Because it will enhance a sense of presence, information on the appearance of the scene before a Mars rover should be presented to its earth operators stereoscopically. A small room can be built with walls that are stereo displays, refreshed by disc or drum memories.

After the robot has criss-crossed an area several times, it will have sent more information to earth than can be displayed at one time. A computer can be used in the manner shown in Fig. 26 to picture what is known about the scene which the robot will encounter if it makes a new traverse across the area.

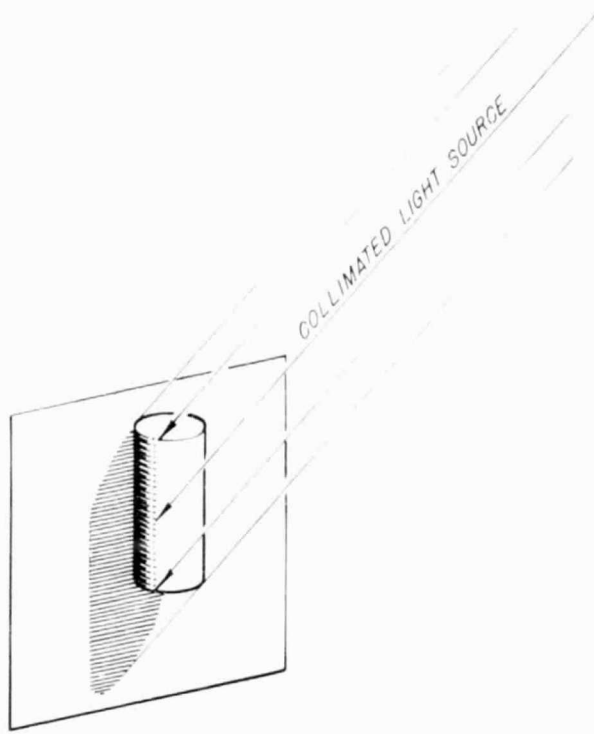


Fig. 25. Effect of a source of collimated light such as sunlight.

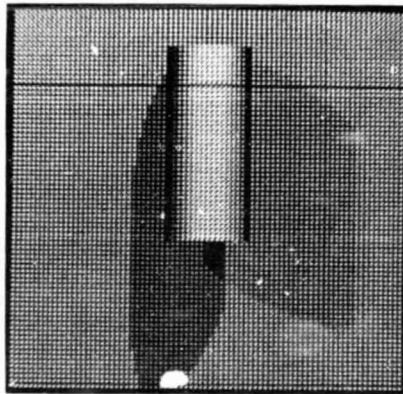


Fig. 26. Reconstruction of the appearance of an object from its shape, its reflectance and the sources of light.

V. STEREO TV CAMERAS

In pursuit of the objectives stated in the Introduction, four families of stereo TV cameras were designed. After two studies, which we called "A" and "B", we designated the cameras as follows:

- C. Single TV camera with two optical paths provided by mirrors (Figs. 0 and 27).
- D. Two TV cameras gimballed separately for vergence, together for pitch. Lenses are of one focal length (Fig. 28).
- E. Two TV cameras with the same gimbaling as in D plus roll and azimuth gimbals for the whole assembly (Fig. 29). Light entering each camera is split into two paths, one of which passes through a long focal length lens, the other through a short focal length lens. Each path of light forms a separate image on the face of the camera tube.

F. Stereo facsimile camera for a crawling vehicle (Ref. 31). The C1 was a conventional TV camera with commercially available stereoc attachment. With its short focal length lenses and 63mm (2.5 in.) interocular distance it proved to be of little value for our work. The C2 was a study. The C3 was built and is now operating.

The first configuration of the C3 was the C3a, shown in Fig. 0, which employed a three-color wheel between the lens and mirror on each side of the system to permit detecting color differences as well as luminance differences in the scene. Each mirror of the C3a was set and bolted in a fixed position. Experiment showed that the angles of the mirrors needed to be adjustable. Accordingly, we mounted each mirror in a bearing and arranged that each mirror be turned by a micrometer acting against a spring (see Fig. 27). We had to remove the color wheels to make room for these mechanisms. In both the Types C3a and C3b, the focal length of the lenses is 50mm and the interocular distance is 21cm (8.25 in.). Focussing is by a

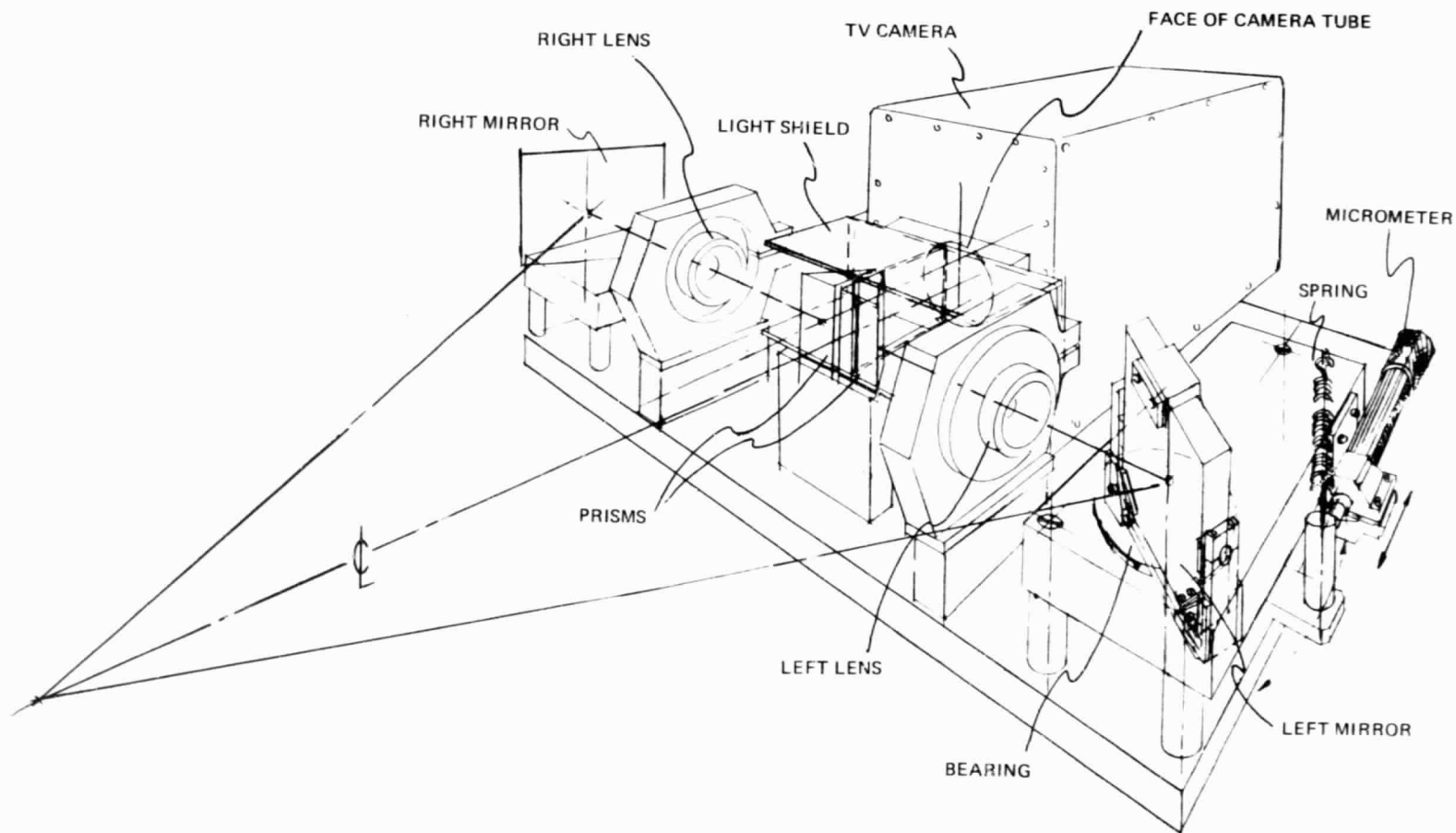


Fig. 27. Type C3b stereo TV camera assembly.

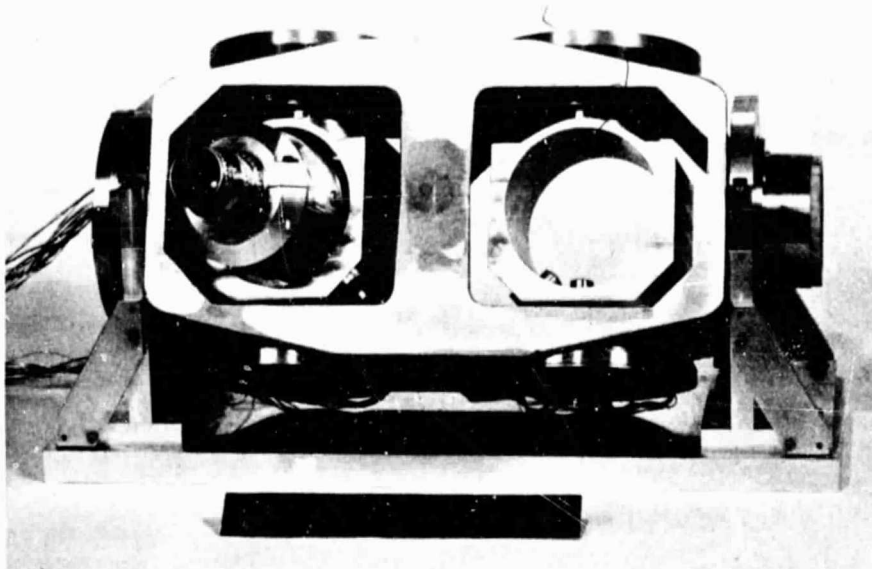


Fig. 28. Type D1 stereo TV camera assembly.
Scale in front of assembly is 6 inches.

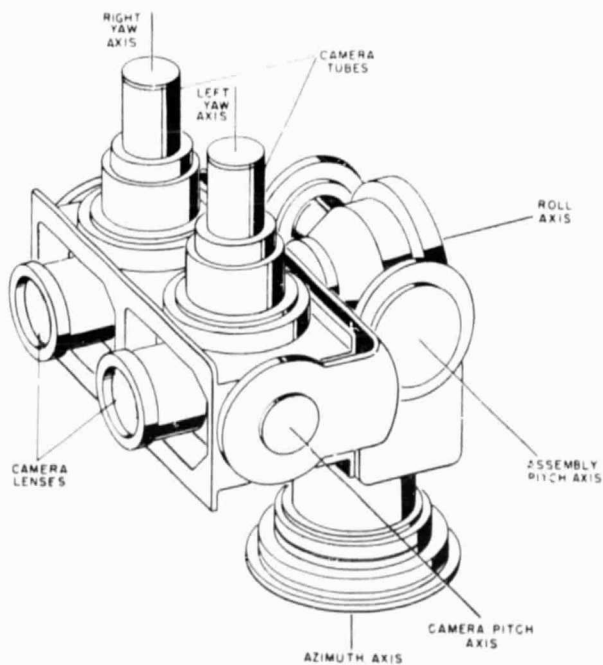


Fig. 29. Type E1 stereo TV camera assembly.

worm gear that moves the TV camera. The C3b could be used in tests like that in II L when adjustments of the camera assembly have been completed and the program STEREO has been modified to compute range from images received along converged axes. The axes need to be converged because the angles of acceptance of the lenses in the Type C3b are smaller than those of the lenses used in the test of II L and the binocular base wider. The advantage of the C3b camera assembly over the assemblies about to be described is that it places both images on the face of one camera tube, thus making it easier to eliminate vertical disparity.

Both the D and E types of stereo TV camera assemblies include two separately gimbaled cameras. The D1 frame has been built and its pitch and yaw electromechanisms operated under servo control (Fig. 27). The E1 assembly, pictured in Fig. 29, is a design on paper of an assembly in which each camera contains optic trains of both 40mm and 400mm focal lengths. The axes of the bearings supporting the cameras are 21cm (8.25 in.) apart in the Type D1, 17.73cm (7 in.) apart in the Type E1 (Ref. 32).

Each optic train of the E1 assembly reflects light upward onto the face of a vertical camera tube, providing the two images shown in Fig. 30. The optic trains are folded upward to keep the front-to-back measurement of the camera as small as possible. The two pitch axes of the assembly will permit it to look both straight down and straight up. The azimuth gimbal will permit it to look in any direction. The roll gimbal will permit it to keep the camera pitch axis horizontal.

The assembly of Fig. 29 has been estimated to weigh 11.4kg (25 lbs.) when made of light-weight spacecraft materials, as much as 34.2kg (75 lbs.) when made of aluminum and steel. In either case the assembly can be mounted most effectively on a rover directly above the axle, as shown in Fig. 31. An arm and one-fingered hand are shown attached to a shoulder of the robot to test the estimate,

made by the visual system, of the size and shape of an object. A second arm with a hand for picking up small objects was also designed but is not shown here.

The human eyes with their narrow-angle high-resolution central fields and their wide-angle low-resolution peripheral fields meet the requirements stated in II N. However, there is no camera tube that can provide, with a short focal-length lens, the angular resolution of human central vision. The only way to achieve this high resolution today is to use a long focal-length lens with a currently available camera tube. As in the human eye each long focal-length lens should be rigidly attached to its finder lens so that when a feature, found in the finder, is centered, it will be in the field of the long focal-length lens. Such a co-axial-input design was achieved in the Type E1 assembly, but no provision was made for focussing. A redesign that provides for focussing is being proposed.

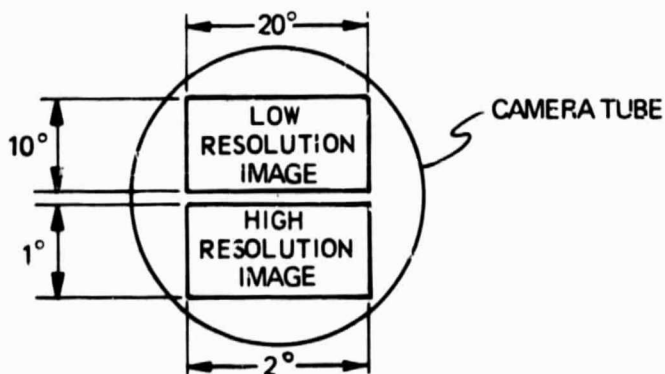


Fig. 30. Diagram of images on the face of each camera tube in the type E1 camera assembly.

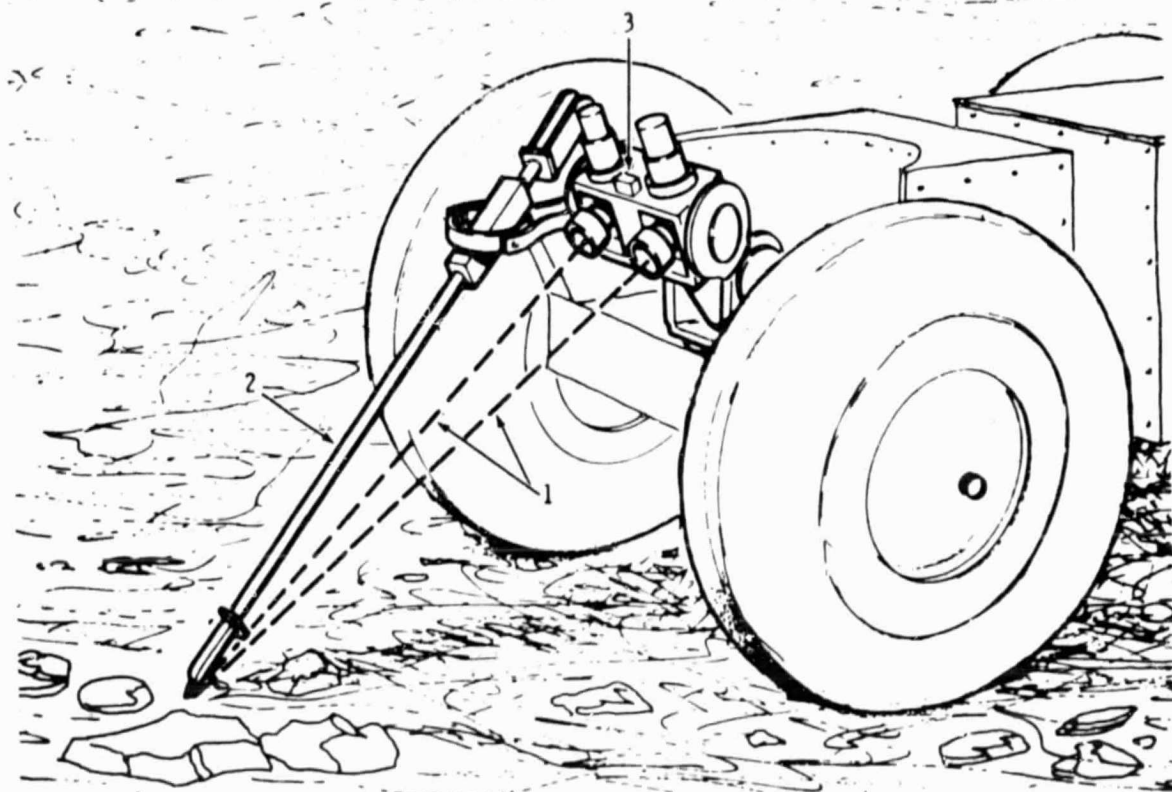
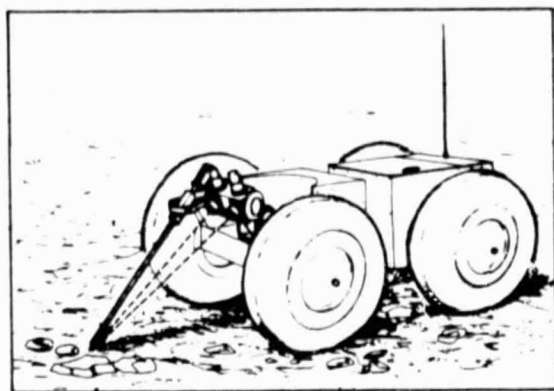


Fig. 31. Possible configuration of a Mars rover. Stereo TV camera views scene along lines of sight (1), while hand and arm (2) feel and accelerometers (3) detect inclination.

VI. SUMMARY AND CONCLUSION

Progress toward automatic recognition of three-dimensional objects, reported here, can be judged from at least two points of view. One is in terms of hardware built and programs operated. A second point of view is of the modelling of the vision of animals that recognize three-dimensional objects.

In this section we summarize our work from the first point of view and draw conclusions from it. The same work seen from the second point of view is summarized in Ref. 16. There it is suggested that the advantage of binocular or stereo vision in a robot, as in an animal, is economy in the computation of form. When the form of a three-dimensional object has been determined, recognition appears achievable by a serial matching of stored features with those detected in the environment. Such serial matching has been demonstrated to be characteristic of human object recognition.

In this report we considered two approaches to the computation of range from a binocular, or stereo, input. In the first approach, described in Section II, gray values in the left image of a scene were matched with gray values in the right image to determine the range of the objects imaged. In the second approach (Section III), edges were extracted from each image.

When using the first approach, we say that an $N \times N$ area of pixels in the left image is "matched" to an $N \times N$ area of pixels in the right image when a preselected percentage of the first set of pixels has gray values that are the same, within a tolerance ϵ , as the gray values of the second set. If the axes of the cameras providing the two images are parallel, the disparity of a match between left and right images is zero for a point at infinity, 36 in our test system for the nearest point. Range is computed from disparity. Our two main problems were, first, how to design this robot vision so that it will reject "spurious" matches between areas of the left

and right views and, second, how to enable it to interpret part of a scene that is ambiguous because only one camera can view it. The method we devised of determining whether or not a match is spurious is to assume that every surface is opaque and that it therefore hides different areas of the background from each camera. By forming left-view and right-view match spaces and comparing them, spurious matches are eliminated.

This first approach was favored in the work reported here because Julesz had performed experiments in which disparity had been computed this way automatically. We repeated Julesz' experiments with random gray-value dot patterns instead of the random gray-value dot patterns he used. We defined a "match space" in which we employed Julesz' methods of both removing spurious matches and of filling ambiguous regions. To demonstrate the results of this processing we devised the range map. Employing next, random square patterns we devised means of eliminating spurious surfaces and of reducing parallelpipeds of matches to planes of matches. Finally, we aimed a TV camera at a Mars-like scene from two positions, with the axes parallel between positions, fed the output of the camera through an analog-to-digital computer into a computer where the images from the camera in its two positions were compared, line by line. The result of matching, eliminating spurious matches and eliminating ambiguities was again a range map.

These steps, when followed by recognition of form, are models of what Julesz calls local and global stereopsis. To take the next step beyond the recognition of form, namely, the automatic recognition of objects, it appears that the second approach is needed, namely, one in which edges are detected and localized in space. In higher vertebrate animals and in our second approach edges are detected prior to comparison between left and right views. In some

cases, however, it may be more efficient to detect edges after comparison of left and right views, for example, by searching the range map.

So that the above operations can be performed rapidly, we have devised a means of shifting as many lines of the images at a time as are needed in computation past a single time-shared computing element. Matched points or edges can be plotted in a single match space equipped with two sets of wiring to provide the two views required of this space for the removal of ambiguities.

So that a stereo TV camera assembly can make a sequence of observations of a scene, we have designed camera assemblies variable in vergence,* pitch, roll and yaw. We have built the D1 assembly that has these properties. A stereo camera assembly, we conclude, must employ both short and long focal length lenses in each camera, the short to provide the finder lenses, the long to provide for identification of features. The one stereo TV camera assembly we have both built and operated (Type C3b) has only short focal-length lenses and is adjustable only in focus and vergence.

We have derived formulas for range accuracy and employed them to predict the effects of different focal lengths, interocular distances and sensor arrays.

Features on the surfaces of objects and spatial relations among these features, transmitted by a robot on Mars to an earth station, can be reconstructed there into the appearance of objects. We have constructed, from such data, the appearance of a cylinder in sunlight.

* Vergence denotes the convergence-divergence adjustment of the two optic axes (Ref. 33).

APPENDIX A

EQUATIONS FOR RANGE AND UNCERTAINTY IN RANGE

A.1 The Geometry of Stereo TV Optics with Parallel Axes

Figure A-1 is a redrawing of Fig. 6a to show the uncertainty in range measurement $\pm \Delta z$ corresponding to the uncertainty of point S on the face of the camera tube. Positive uncertainty is $+\Delta z$ or PM, negative uncertainty $-\Delta z$ or PR. The angle of uncertainty δ is also introduced in preparation for a discussion of the first strategy of II B. In addition to the axis z, there is a parallel axis q in the x-z plane, which bisects the interocular distance, 2b. The range, z, or P is measured from the optical centers of the lenses, L, which are assumed to behave like pinhole lenses.

From the similarity of triangles POL and LAS in Fig. A.1,

$$\begin{aligned}\frac{z}{b} &= \frac{f}{s} \\ z &= \frac{bf}{s}\end{aligned}\tag{A-1}$$

but in Fig. 6,

$$\begin{aligned}s &= d_L = d_R = \frac{d}{2} \\ \therefore z &= \frac{bf}{d/2} = \frac{2bf}{d}\end{aligned}\tag{1) on p. 19}$$

Thus, for a system with parallel optical axes where range is measured from the optical centers of the lenses, the nominal value of the range is the product of the interocular distance and the focal length, divided by the disparity between the positions of the image on the image plane.

A.2 Derivation of Equation of Range Uncertainty for Stereo TV Cameras with Parallel Axes

If the uncertainty of the position S on the image plane is either Δs or $-\Delta s$, then the uncertainty in the range is Δz or $-\Delta z$, respectively.

From Fig. A-1,

$$\frac{f}{s - \Delta s} = \frac{z + \Delta z}{b} \quad (\text{A-2})$$

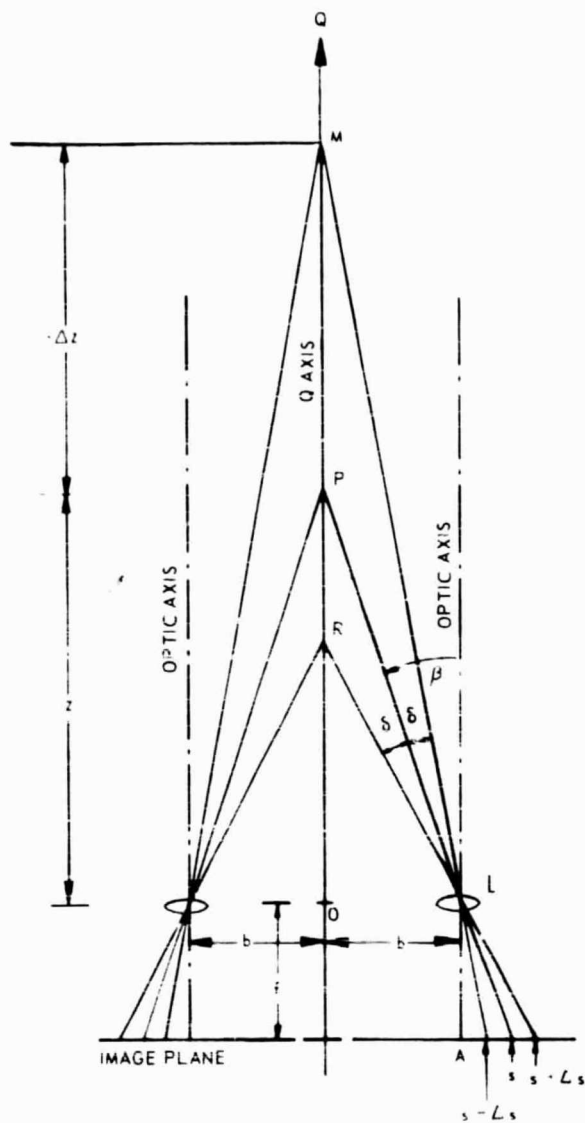
$$z + \Delta z = \frac{bf}{s - \Delta s} \quad (\text{A-3})$$

Subtracting Eq (A-1) from (A-3), we obtain the range uncertainty Δz due to uncertainty in a position S on the image plane (or equivalently an angular uncertainty δ in determining β):

$$\Delta z = \frac{bf(s - \Delta s) - bfs}{s(s - \Delta s)} = \frac{bf \cdot \Delta s}{s(s - \Delta s)} \quad (\text{A-4})$$

Rearranging Eq. (A-1) and substituting into Eq. (A-4),

$$\Delta z = \frac{\Delta s \cdot z^2}{bf - \Delta s \cdot z} \quad (2) \text{ on page 38}$$



SYMBOL DEFINITIONS

- z = RANGE OF A POINT P FROM O
- A = POINT WHERE OPTIC AXIS PIERCES IMAGE PLANE
- s = DISTANCE ON THE IMAGE PLANE FROM POINT A TO IMAGE OF POINT P
- $\Delta z = PM$ = UNCERTAINTY IN RANGE CORRESPONDING TO UNCERTAINTY IN THE POSITION OF POINT S, Δs
- $-\Delta z = PR$

Fig. A-1. Geometry of parallel optics for range finding.

REFERENCES

1. Sutro, L. L., D. B. Moulton, et al., "1963 Advanced Sensor Investigations", Report R-470, Charles Stark Draper Laboratory, MIT, 1964, pp. 1-35.
2. Sutro, L. L., "Advanced Sensor and Control System Studies, 1964 to September 1965", Report R-519, Charles Stark Draper Laboratory, MIT, 1966, pp. 16-33.
3. Moreno-Diaz, R., "An Analytical Model of the Group 2 Ganglion Cell in the Frog's Retina", Report E-1858, Charles Stark Draper Laboratory, MIT, 1965.
4. Moreno-Diaz, R., "An Analytical Model of the 'Bug Detector' Ganglion Cell in the Frog's Retina", Cybernetic Problems in Bionics, H. L. Oestreicher and D. R. Moore, Eds., Gordon and Breach, N. Y., 1968, pp. 481-491.
5. Sutro, L. L., "Information Processing and Data Compression for Exobiology Missions", Advances in Astronautical Science, Vol. 22, The Search for Extraterrestrial Life, Scholarly Publications, Inc., Sun Valley, Cal., 1967, pp. 347-378, and Report R-545, Charles Stark Draper Laboratory, MIT, 1966.
6. Sutro, L. L., "Proposed Electronics to Represent the Properties of a Frog's Eye", Cybernetic Problems in Bionics, H. L. Oestreicher and D. R. Moore, Eds., Gordon and Breach, N. Y., 1968, pp. 811-819.
7. Sutro, L. L., "First Steps Toward a Model of The Vertebrate Central Nervous System", Cybernetics, Artificial Intelligence and Ecology, Proceedings of the Fourth Annual Symposium of the American Society for Cybernetics, H. W. Robinson and D. E. Knight, Eds., Spartan Books, N. Y., 1972, pp. 225-252.

8. Lettvin, J. L., H. R. Maturana, W. S. McCulloch and W. H. Pitts, "What a Frog's Eye Tells a Frog's Brain", Proceedings of the L. R. E., Vol. 47, No. 11, November, 1959.
9. Sutro, L. L., et al., "Development of Visual, Contact and Decision Subsystems for a Mars Rover", Report R-565, Charles Stark Draper Laboratory, MIT, 1967, pp. 15-16 (out of print).
10. Fukishima, K., "Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements", IEEE Transactions on System Science and Cybernetics, Vol. SSC-5, No. 4 (Oct., 1969), pp. 332-333.
11. Fukishima, K., "A Feature Extractor for Curvilinear Patterns: A Design Suggested by the Mammalian Visual System", Kybernetik 7 (1970), pp. 153-160.
12. Sutro, L. L. and J. B. Lerman, "Robot Vision", Remotely Manned Systems in Space, Proceedings of First National Conference, edited by Ewald Heer, California Institute of Technology, Pasadena, Cal., 1973.
13. Held, R., et al., "Locating and Identifying: Two Modes of Visual Processing. A Symposium", Psychologische Forschung 31, 42/43 (1967).
14. Julesz, B., Foundations of Cyclopean Perception, The University of Chicago Press, Chicago, 1971, p. 144.
15. Sutro, L. L., and W. L. Kilmer, "An Assembly of Computers to Command and Control a Robot", first published in the Proceedings of the 1969 Spring Joint Computer Conference, AFIPS Press, Montvale, N.J., pp. 113-137. This paper was revised and republished both in ANALOG-Science Fiction, Science Fact, May 1970, and as Report R-582-1, Charles Stark Draper Laboratory, MIT, 1969.

16. Sutro, L. L. and W. L. Kilmer, "Development of Decision-Making Devices Inspired by the Animal Nervous System," Report R-636, Charles Stark Draper Laboratory, MIT, Cambridge, Mass. (in preparation).
17. Sutro, L. L. and W. L. Kilmer, "Development of Decision-Making Network Based Upon a Model of the Reticular Formation" Report R-736, Charles Stark Draper Laboratory, MIT, Cambridge, Mass. (in preparation).
18. Sutro, L. L. and W. L. Kilmer, op. cit., Ref. 15, p. 131. Also in Report R-582-1, Charles Stark Draper Laboratory, MIT, 1969, p. 52.
19. Julesz, B., "Towards the Automation of Binocular Depth Perception", Proceedings of the IFIP Congress, 1962, North Holland Publishing Co., Amsterdam, 1963, pp. 439-443.
20. Lerman, J. B., "Computer Processing of Stereo Images for the Automatic Extraction of Range", thesis submitted in partial fulfillment for the degrees of bachelor of science and master of science, Department of Electrical Engineering, MIT, Report T-531, Charles Stark Draper Laboratory, MIT, 1970 (out of print).
21. Nitzan, D., "Stereopsis Error Analysis", Technical Note 71, Artificial Intelligence Center, Stanford Research Institute, Menlo Park, California, 1972.
22. Julesz, B., op. cit., Ref. 14, p. 150.
23. Pettigrew, J. D., "The Neurophysiology of Binocular Vision", Scientific American, Vol. 227, No. 2, August, 1972, pp. 84-96.
24. Julesz, B., op. cit., Ref. 14, p. 1.

25. Julesz, B. , "Texture and Visual Perception", Scientific American, Vol. 212, No. 2, February 1965, p. 46.
26. Sutro, L. L. , Notes of Conversations with W.S. McCulloch, 1958-1969.
27. Prewitt, J. , M. S. , "Object Enhancement and Extraction", in Picture Processing and Psychopictorics, edited by B. S. Lipkin and A. Rosenfeld, Academic Press, New York, 1970, pp. 75-149.
28. Rosenfeld, A. , "A Nonlinear Edge Detection Technique", Proceedings of the I.E.E.E. , May 1970, pp. 814-816.
29. Sutro, L. L. and W. L. Kilmer, op. cit., Ref. 15, pp. 123 and 124. Also in Report R-582-1, Charles Stark Draper Laboratory, MIT, 1969, pp. 28-31.
30. Sutro, L. L. , "A Model of Visual Space", Biological Prototypes and Synthetic Systems, Vol. 1, E.E. Bernard and M. R. Kare, eds. , Plenum Press, Inc. , New York, 1962, pp. 75-87. In reading this paper, substitute the word "luminance" for the word "brightness".
31. Sutro, L. L. and W. L. Kilmer, op. cit., Ref. 15, p. 130. Also in Report R-582-1, Charles Stark Draper Laboratory, MIT, 1969, pp. 47 and 48.
32. Sutro, L. L. , W. L. Kilmer, R. Moreno-Diaz, W.S. McCulloch et al. , op. cit., Ref. 9, pp. 23-30 and 85-86.
33. Julesz, B. , op. cit. , Ref. 22, p. 52.
34. Ogle, K. N. , "Researches in Binocular Vision", Hafner Publishing Co. , New York, 1964, pp. 166, 173ff and 281.